CHAPTER ELEVEN

Test Bias

Concepts and Criticisms

ARTHUR R. JENSEN

As one who has been reading about test bias now for over 30 years, I have noticed a quite dramatic change in this literature within just the last decade. This development was auspicious, perhaps even essential, for the production of my most recent book, *Bias in Mental Testing* (1980a). Developments in the last decade made it possible to present a fairly comprehensive and systematic treatment of the topic. Prior to the 1970s, the treatment of test bias in the psychological literature was fragmentary, unsystematic, and conceptually confused. Clear and generally agreed-upon definitions of bias were lacking, as was a psychometrically defensible methodology for objectively recognizing test bias. The study of test bias, in fact, had not yet become a full-fledged subject in the field of psychometrics. The subject lacked the carefully thought-out rationale and statistical methodology that psychometrics had long invested in such topics as reliability, validity, and item selection.

All this has changed markedly in recent years. Test bias has now become one of the important topics in psychometrics. It is undergoing the systematic conceptual and methodological development worthy of one of the most technically sophisticated branches of the behavioral sciences. The earlier scattered and inchoate notions about

Parts of this chapter are taken from "Précis of Bias in Mental Testing" by A. R. Jensen, *Behavioral and Brain Sciences*, 1980, *3*, 325–333.

ARTHUR R. JENSEN • Institute of Human Learning, University of California, Berkeley, California 94720.

bias have been sifted, rid of their patent fallacies, conceptualized in objective terms, and operationalized by statistical methods. What is emerging is a theoretical rationale of the nature of test bias, some rather clearly formulated, mutually consistent definitions, and statistically testable criteria of bias. Moreover, a large fund of impressively consistent empirical evidence has been amassed in connection with this discipline, finally permitting objective, often definitive, answers to the long-standing question of racial-cultural bias in many of the standardized mental tests widely used in America today in schools, colleges, and the armed forces, and for job selection.

The editors have asked me to act as a commentator on all the preceding chapters in this volume. Before taking up the many specific points in this task, however, I should first present a succinct overview of the main concepts and findings in this field, as I see it. I have presented it all in much greater detail in *Bias in Mental Testing*.

NATURE OF MENTAL TESTS

Mental ability tests are a means of quantifying individual differences in a variety of capabilities classified as *mental*. *Mental* means only that the individual differences in the capabilities elicited by the test are not primarily the result of differences in sensory acuity or motor dexterity and coordination. *Ability* implies three things: (1) conscious, voluntary behavior; (2) maximum, as contrasted with typical, performance (at the time); and (3) an objective standard for rating performance on each unit or item of the test, such as correct versus incorrect, pass versus fail, or measurement of rate, such as number of test units completed per unit time or average time per unit. By *objective standard* one means that differences in performance on any unit of the test can be judged as "better than" or "worse than" with universal agreement, regardless of possible disagreements concerning the social value or importance that may be placed on the performance.

A mental test is composed of a number of items having these properties, each item affording the opportunity to the person taking the test to demonstrate some mental capability as indicated by his or her objectively rated response to the item. The total raw score on the test is the sum of the ratings (e.g., "pass" versus "fail" coded as 1 and 0) of the person's responses to each item in the test.

The kinds of items that compose a test depend on its purpose and on certain characteristics of the particular population for which its

use is intended, such as age, language, and educational level. The set of items for a particular test is generally devised and selected in accordance with some combination of the following criteria: (1) a psychological theory of the nature of the ability the test is intended to measure; (2) the characteristics of the population for which it is intended; (3) the difficulty level of the items, as indicated by the proportion of the target population who "pass" the item, with the aim of having items that can discriminate between persons at every level of ability in the target population; (4) internal consistency, as indicated by positive intercorrelations among the items making up the test, which means that all the items measure some common factor; and (5) the "item characteristic curve," which is the function relating (a) the probability of an individual's passing a given item to (b) the individual's total score on the test as a whole (if a is not a monotonically increasing function of b, the item is considered defective). The individual items (or their common factors) are then correlated with external performance criteria (e.g., school grades, job performance ratings).

The variety of types of test items in the whole mental abilities domain is tremendous and can scarcely be imagined by persons outside the field of psychological testing. Tests may be administered to groups or individuals. They can be verbal, nonverbal, or performance (i.e., requiring manipulation or construction) tests. Within each of these main categories, there is a practically unlimited variety of item types. The great number of apparently different kinds of tests, however, does not correspond to an equally large number of different, measurable abilities. In other words, a great many of the superficially different tests—even as different as *vocabulary* and *block designs* (constructing designated designs with various colored blocks)—must to some extent measure the same abilities.

General Intelligence or *g*

One of the great discoveries in psychology, originally made by Charles E. Spearman in 1904, is that, in an unselected sample of the general population, *all* mental tests (or test items) show nonzero positive intercorrelations. Spearman interpreted this fact to mean that every mental test measures some ability that is measured by all other mental tests. He labeled this common factor *g* (for "general factors"), and he developed a mathematical technique, known as *factor analysis,* that made it possible to determine (1) the proportion of the total variance (i.e., individual differences) in scores on a large collection of

diverse mental tests that is attributable to individual variation in the general ability factor, or *g*, that is common to all of the tests, and (2) the degree to which each test measures the *g* factor, as indicated by the test's correlation with the *g* factor (termed the test's *factor loading*).

Later developments and applications of factor analysis have shown that in large, diverse collections of tests there are also other factors in addition to *g*. Because these additional factors are common only to certain groups of tests, they are termed *group factors*. Well-established group factors are verbal reasoning, verbal fluency, numerical ability, spatial-perceptual ability, and memory. However, it has proved impossible to devise tests that will measure only a particular group factor without also measuring *g*. All so-called factor-pure tests measure *g* plus some group factor. Usually, considerably more of the variance in scores on such tests is attributable to the *g* factor than to the particular group factor the test is designed to measure. The total score on a test composed of a wide variety of items reflects mostly the *g* factor.

Spearman's principle of the *indifference of the indicator* recognizes the fact that the *g* factor can be measured by an almost unlimited variety of test items and is therefore conceptually independent of the particular form or content of the items, which are merely vehicles for the behavioral manifestations of *g*. Spearman and the psychologists following him identify *g* with general mental ability or general intelligence. It turns out that intelligence tests (henceforth referred to as *IQ tests*), which are judged to be good indicators of intelligence by a variety of criteria other than factor analysis, have especially high *g* loadings when they are factor-analyzed among a large battery of diverse tests.

To gain some insight into the nature of *g*, Spearman and many others have compared literally hundreds of tests and item types in terms of their *g* loadings to determine the characteristics of those items that are the most and the least *g*-loaded. Spearman concluded that *g* is manifested most in items that involve "relation eduction," that is, seeing relationships between elements, grasping concepts, drawing inferences—in short, inductive and deductive reasoning and problem solving. "Abstractness" also enhances an item's *g* loading, such as being able to give the meaning of an abstract noun (e.g., *apotheosis*) as contrasted with a concrete noun (e.g., *aardvark*) when both words are equated for difficulty (i.e, percentage passing in the population). An item's *g* loading is independent of its difficulty. For example, certain tests of rote memory can be made very difficult, but

they have very low *g* loadings. *Inventive* responses to novel situations are more highly *g*-loaded than responses that depend on recall or reproduction of past acquired knowledge or skill. The *g* factor is related to the *complexity* of the mental manipulations or transformations of the problem elements required for solution. As a clear-cut example, forward digit span (i.e., recalling a string of digits in the same order as the input) is less *g*-loaded than backward digit span (recalling the digits in reverse order), which requires more mental manipulation of the input before arriving at the output. What we think of as "reasoning" is a more complex instance of the same thing. Even as simple a form of behavior as *choice reaction time* (speed of reaction to either one or the other of two signals) is more *g*-loaded than is *simple reaction time* (speed of reaction to a single signal). It is a well-established empirical fact that more complex test items, regardless of their specific form or content, are more highly correlated with one another than are less complex items. In general, the size of the correlation between any two tests is directly related to the product of the tests' *g* loadings.

Tests that measure *g* much more than any other factors can be called *intelligence tests.* In fact, *g* accounts for most of the variance not only in IQ tests, but in most of the standardized aptitude tests used by schools, colleges, industry, and the armed services, regardless of the variety of specific labels that are given to these tests. Also, for persons who have been exposed to essentially the same schooling, the general factor in tests of scholastic achievement is very highly correlated with the *g* factor of mental tests in general. This correlation arises not because the mental tests call for the specific academic information or skills that are taught in school, but because the same *g* processes that are evoked by the mental tests also play an important part in scholastic performance.

Is the *g* factor the same ability that the layperson thinks of as "intelligence"? Yes, very largely. Persons whom laypeople generally recognize as being very "bright" and persons recognized as being very "dull" or retarded do, in fact, differ markedly in their scores on tests that are highly *g*-loaded. In fact, the magnitudes of the differences between such persons on various tests are more closely related to the tests' *g* loadings than to any other characteristics of the tests.

The practical importance of *g*, which is measured with useful accuracy by standard IQ tests, is evidenced by its substantial correlations with a host of educationally, occupationally, and socially valued variables. The fact that scores on IQ tests reflect something more profound than merely the specific knowledge and skills acquired in

school or at home is shown by the correlation of IQ with brain size (Van Valen, 1974), the speed and amplitude of evoked brain potentials (Callaway, 1975), and reaction times to simple lights or tones (Jensen, 1980b).

CRITICISM OF TESTS AS CULTURALLY BIASED

Because IQ tests and other highly g-loaded tests, such as scholastic aptitude and college entrance tests and many employment selection tests, show sizable average differences between majority and minority (particularly black and Hispanic) groups, and between socioeconomic classes, critics of the tests have claimed that the tests are culturally biased in favor of the white middle class and against certain racial and ethnic minorities and the poor. Asians (Chinese and Japanese) rarely figure in these claims, because their test scores, as well as their performance on the criteria the tests are intended to predict, are generally on a par with those of the white population.

Most of the attacks on tests, and most of the empirical research on group differences, have concerned the observed average difference in performance between blacks and whites on virtually all tests of cognitive ability, amounting to about one standard deviation (the equivalent of 15 IQ points). Because the distribution of IQs (or other test scores) approximately conforms to the normal or bell-shaped curve in both the white and the black populations, a difference of one standard deviation between the means of the two distributions has quite drastic consequences in terms of the proportions of each population that fall in the upper and lower extremes of the ability scale. For example, an IQ of about 115 or above is needed for success in most highly selective colleges; about 16% of the white as compared with less than 3% of the black population have IQs above 115, that is, a ratio of about 5 to 1. At the lower end of the IQ distribution, IQs below 70 are generally indicative of mental retardation: Anyone with an IQ below 70 is seriously handicapped, educationally and occupationally, in our present society. The percentage of blacks with IQs below 70 is about six times greater than the percentage of whites. Hence blacks are disproportionately underrepresented in special classes for the academically "gifted," in selective colleges, and in occupations requiring high levels of education or of mental ability, and they are seen in higher proportions in classes for "slow learners" or the "educable mentally retarded." It is over such issues that tests, or the uses of tests in schools, are literally on trial, as in the well-

known *Larry P.* case in California, which resulted in a judge's ruling that IQ tests cannot be given to blacks as a basis for placement in special classes for the retarded. The ostensible justification for this decision was that the IQ tests, such as the Stanford–Binet and the Wechsler Intelligence Scale for Children, are culturally biased.

The claims of test bias, and the serious possible consequences of bias, are of great concern to researchers in psychometrics and to all psychologists and educators who use tests. Therefore, in *Bias in Mental Testing,* I have tried to do essentially three things: (1) to establish some clear and theoretically defensible definitions of test bias, so we will know precisely what we are talking about; (2) to explicate a number of objective, operational psychometric criteria of bias and the statistical methods for detecting these types of bias in test data; and (3) to examine the results of applying these objective criteria and analytic methods to a number of the most widely used standardized tests in school, college, the armed services, and civilian employment.

TEST SCORES AS PHENOTYPES

Let me emphasize that the study of test bias *per se* does not concern the so-called nature–nurture or heredity–environment issue. Psychometricians are concerned with tests only as a means of measuring *phenotypes.* Test scores are treated as such a means. Considerations of their validity and their possible susceptibility to biases of various kinds in all of the legitimate purposes for which tests are used involve only the phenotypes. The question of the correlation between test scores (i.e., the phenotypes) and genotypes is an entirely separate issue in quantitative genetics, which need not be resolved in order for us to examine test bias at the level of psychometrics. It is granted that individual differences in human traits are a complex product of genetic and environmental influences; this product constitutes the *phenotype.* The study of test bias is concerned with bias in the measurement of phenotypes and with whether the measurements for certain classes of persons are systematically distorted by artifacts in the tests or testing procedures. Psychometrics as such is *not* concerned with estimating persons' genotypes from measurements of their phenotypes and therefore does not deal with the question of possible bias in the estimation of genotypes. When we give a student a college aptitude test, for example, we are interested in accurately assessing his or her level of developed ability for doing college work, because it is the student's developed ability that actually pre-

dicts his or her future success in college, and not some hypothetical estimate of what his or her ability *might* have been if he or she had grown up in different circumstances.

The scientific explanation of racial differences in measurements of ability, of course, must examine the possibility of test bias *per se*. If bias is not found, or if it is eliminated from particular tests, and a racial difference remains, then bias is ruled out as an adequate explanation. But no other particular explanations, genetic or environmental, are thereby proved or disproved.

MISCONCEPTIONS OF TEST BIAS

There are three popular misconceptions or fallacies of test bias that can be dismissed on purely logical grounds. Yet, they have all figured prominently in public debates and court trials over the testing of minorities.

Egalitarian Fallacy

This fallacy holds that any test that shows a mean difference between population groups (e.g., races, social class, sexes) is therefore necessarily biased. Men measure taller than women; therefore yardsticks are sexually biased measures of height. The fallacy, of course, is the unwarranted *a priori* assumption that all groups are equal in whatever the test purports to measure. The converse of this fallacy is the inference that the *absence* of a mean difference between groups indicates that the test is unbiased. It could be that the test bias is such as to equalize the means of groups that are truly unequal in the trait the test purports to measure. As scientifically egregious as this fallacy is, it is interesting that it has been invoked in most legal cases and court rulings involving tests.

Culture-Bound Fallacy

This fallacy is the mistaken belief that because test items have some cultural content they are necessarily culture-biased. The fallacy is in confusing two distinct concepts: *culture loading* and *culture bias*. (Culture-*bound* is a synonym for *culture-loaded*.) These terms do not mean the same thing.

Tests and test items can be ordered along a continuum of culture loading, which is the specificity or generality of the informational

content of the test items. The narrower or less general the culture in which the test's information content could be acquired, the more culture-loaded it is. This can often be roughly determined simply by inspection of the test items. A test item requiring the respondent to name three parks in Manhattan is more culture-loaded than the question "How many 20-cents candy bars can you buy for $1?" To the extent that a test contains cultural content that is generally peculiar to the members of one group but not to the members of another group, it is liable to be culture-biased with respect to comparisons of the test scores between the groups or with respect to predictions based on their test scores.

Whether the particular cultural content actually causes the test to be biased with respect to the performance of any two (or more) groups is a separate issue. It is an empirical question. It cannot be answered merely by inspection of the items or subjective impressions. A number of studies have shown that although there is a high degree of agreement among persons (both black and white) when they are asked to judge which test items appear the most and the least *culture loaded,* persons can do no better than chance when asked to pick out the items that they judge will discriminate the most or the least between any two groups, say, blacks and whites. Judgments of *culture loading* do not correspond to the actual population discriminability of items. Interestingly, the test items most frequently held up to ridicule for being "biased" against blacks have been shown by empirical studies to discriminate less between blacks and whites than the average run of items composing the tests! Items judged as "most culture-loaded" have not been found to discriminate more between whites and blacks than items judged as "least culture loaded." In fact, one excellently designed large-scale study of this matter found that the average white–black difference is *greater* on the items judged as "least cultural" than on items judged "most cultural," and this remains true when the "most" and "least" cultural items are equated for difficulty (percentage passing) in the white population (McGurk, 1967).

Standardization Fallacy

This fallacy is the belief that a test that was constructed by a member of a particular racial or cultural population and standardized or "normed" on a representative sample of that same population is therefore necessarily biased against persons from all other populations. This conclusion does not logically follow from the

premises, and besides, the standardization fallacy has been empirically refuted. For example, representative samples of Japanese (in Japan) average about 6 IQ points higher than the American norms on the performance scales (nonverbal) of the Wechsler Intelligence Test, which was constructed by David Wechsler, an American psychologist, and standardized in the U.S. population. Arctic Eskimos score on a par with British norms on the Progressive Matrices Test, devised by the English psychologist J. C. Raven and standardized in England and Scotland.

THE MEANING OF BIAS

There is no such thing as test bias in the abstract. Bias must involve a specific test used in two (or more) specific populations.

Bias means *systematic* errors of measurement. All measurements are subject to *random* errors of measurement, a fact that is expressed in terms of the coefficient of *reliability* (i.e., the proportion of measurement) and the *standard error of measurement* (i.e., the standard deviation of random errors). *Bias,* or systematic error, means that an obtained measurement (test score) consistently *over*estimates (or *under*estimates) the true (error-free) value of the measurement for members of one group as compared with members of another group. In other words, a biased test is one that yields scores that have a different meaning for members of one group from their meaning for members of another. If we use an elastic tape measure to determine the heights of men and women, and if we stretch the tape every time we measure a man but do not stretch it whenever we measure a woman, the obtained measurements will be biased with respect to the sexes; a man who measures 5′6″ under those conditions may actually be seen to be half a head taller than a woman who measures 5′6″, when they stand back to back. There is no such direct and obvious way to detect bias in mental tests. However, there are many indirect indicators of test bias.

Most of the indicators of test bias are logically one-sided or nonsymmetrical; that is, statistical significance of the indicator can demonstrate that bias exists, but nonsignificance does not assure the absence of bias. This is essentially the well-known statistical axiom that it is impossible to prove the null hypothesis. We can only reject it. Unless a test can be shown to be biased at some acceptable level of statistical significance, it is presumed to be unbiased. The more diverse the possible indicators of bias that a test "passes" without sta-

tistical rejection of the null hypothesis (i.e., "no bias"), the stronger is the presumption that the test is unbiased. Thus, in terms of statistical logic, the burden of proof is on those who claim that a test is biased.

The consequences of detecting statistically significant bias for the practical use of the test is a separate issue. They will depend on the actual magnitude of the bias (which can be trivial, yet statistically significant) and on whether the amount of bias can be accurately determined, thereby permitting test scores (or predictions from scores) to be corrected for bias. They will also depend on the availability of other valid means of assessment that could replace the test and are *less* biased.

EXTERNAL AND INTERNAL MANIFESTATIONS OF BIAS

Bias is suggested, in general, when a test behaves differently in two groups with respect to certain statistical and psychometric features which are conceptually independent of the distributions of scores in the two populations. Differences between the score distributions, particularly between measures of central tendency, cannot themselves be criteria of bias, as these distributional differences are the very point in question. Other objective indicators of bias are required. We can hypothesize various ways that our test statistics should differ between two groups if the test were in fact biased. These hypothesized psychometric differences must be independent of distributional differences in test scores, or they will lead us into the egalitarian fallacy, which claims bias on the grounds of a group difference in central tendency.

Appropriate indicators of bias can be classified as *external* and *internal*.

External Indicators

External indicators are correlations between the test scores and other variables external to the test. An unbiased test should show similar correlations with other variables in the two or more populations. A test's *predictive validity* (the correlation between test scores and measures of the criterion, such as school grades or ratings of job performance) is the most crucial external indicator of bias. A significant group difference in validity coefficients would indicate bias. Of course, statistical artifacts that can cause spurious differences in correlation (or validity) coefficients must be ruled out or cor-

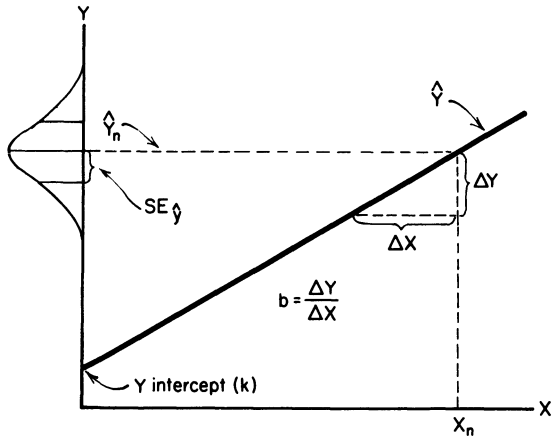

FIGURE 1. Graphic representation of the regression of criterion measurements (Y) on test scores (X), showing the *slope* (b) of the regression line $\hat{Y}$, the Y intercept (k), and the *standard error of estimate* ($SE_{\hat{Y}}$). A test score X_n would have a predicted criterion performance of $\hat{Y}_n$ with a standard error of $SE_{\hat{Y}}$. The regression line $\hat{Y}$ yields the statistically best prediction of the criterion Y for any given value of X. Biased prediction results if one and the same regression line is used to predict the criterion performance of individuals in majority and minority groups when, in fact, the regression lines of the separate groups differ significantly in intercepts, slopes, or standard errors of estimate. The test will yield unbiased predictions for all persons regardless of their group membership if these regression parameters are the same for every group.

rected—such factors as restriction of the "range of talent" in one group, floor or ceiling effects on the score distributions, and unequal reliability coefficients (which are *internal* indicators of bias). Also, the intercept and slope of the regression of criterion measures on test scores, and the standard error of estimate, should be the same in both populations for an unbiased test. The features of the regression of criterion measurements (Y) on test scores (X) are illustrated in Figure 1.

Another external indicator is the correlation of raw scores with age, during the period of mental growth from early childhood to maturity. If the raw scores reflect degree of mental maturity, as is claimed for intelligence tests, then they should show the same correlation with chronological age in the two populations. A significant difference in correlations, after ruling out statistical artifacts, would indicate that the test scores have different meanings in the two groups. Various kinship correlations (e.g., monozygotic and dizygotic

twins, full siblings, and parent–child) should be the same in different groups for an unbiased test.

Internal Indicators

Internal indicators are psychometric features of the test data themselves, such as the test's internal consistency reliability (a function of the interitem correlations), the factorial structure of the test or a battery of subtests (as shown by factor analysis), the rank order of item difficulties (percentage passing each item), the significance and magnitude of the items × groups interaction in the analysis of variance of the item matrix for the two groups (see Figure 2), and the relative "pulling power" of the several error "distractors" (i.e., response alternatives besides the correct answer) in multiple-choice test items. Each of these psychometric indicators is capable of revealing statistically significant differences between groups, if such differences exist. Such findings would indicate bias, on the hypothesis that

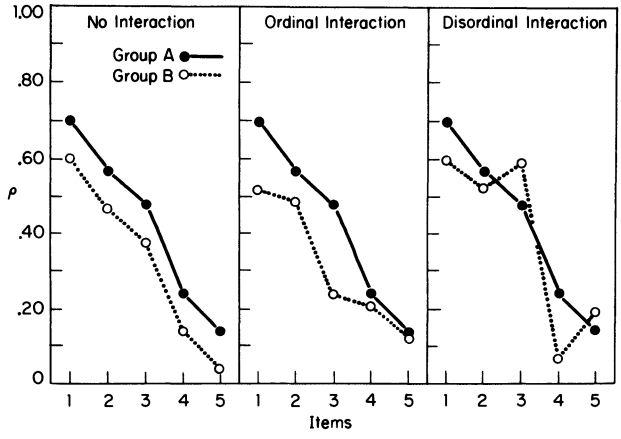

FIGURE 2. Graphic representation of types of items × groups interaction for an imaginary five-item test. Item difficulty (proportion passing the item) is shown on the ordinate; the five items are shown on the baseline. When the item difficulties for two groups, A and B, are perfectly parallel, there is no interaction. In *ordinal* interaction, the item difficulties of Groups A and B are not parallel but maintain the same rank order. In *disordinal interaction*, the item difficulties have a different rank order in the two groups. Both types of interaction are detectable by means of correlational analysis and analysis of variance of the item matrix. Significant items × groups interactions are internal indicators of test bias; that is, such interactions reveal that the test items do not show the same relative difficulties for both groups.

these essential psychometric features of tests should not differ between populations for an unbiased test.

Undetectable Bias

Theoretically, there is a type of bias that could not be detected by any one or any combination of these proposed external and internal indicators of bias. It would be a *constant* degree of bias for one group that affects every single item of a test equally, thereby depressing all test scores in the disfavored group by a constant amount; and the bias would have to manifest the same relative effects on *all* of the external correlates of the test scores. The bias, in effect, would amount to subtracting a constant from every unit of measured performance in the test, no matter how diverse the units, and subtracting a constant from the test's external correlates for the disfavored group. No model of culture bias has postulated such a uniformly pervasive influence. In any case, such a uniformly pervasive bias would make no difference to the validity of tests for any of their usual and legitimate uses. Such an *ad hoc* hypothetical form of bias, which is defined solely by the impossibility of its being empirically detected, has no scientific value.

BIAS AND UNFAIRNESS

It is essential to distinguish between the concepts of *bias* and *unfairness*. Bias is an objective, statistical property of a test in relation to two or more groups. The concept of *unfairness* versus the *fair* use of tests refers to the way that tests are used and implies a philosophic or value judgment concerning procedures for the educational and employment selection of majority and minority groups. The distinction between bias and unfairness is important, because an unbiased test may be used in ways that can be regarded as fair or unfair in terms of one's philosophic position regarding selection strategies, for example, in the question of "color-blind" versus preferential or quota selection of minorities. A statistically biased test can also be used either fairly or unfairly. If one's selection philosophy permits identification of each individual's group membership, then a biased test can often be used fairly for selection, for example, by using separate (but equally effective) regression equations for majority and minority persons in predicting criterion performance, or by entering group

membership (in addition to test scores) in the regression equation to predict future performance.

EMPIRICAL EVIDENCE ON EXTERNAL INDICATORS OF BIAS

The conclusions based on a preponderance of the evidence from virtually all of the published studies on each of the following external criteria of bias are here summarized for all tests that can be regarded as measures of general ability, such as IQ tests, scholastic aptitude, and "general classification" tests. This excludes only very narrow tests of highly specialized skills or aptitudes that have relatively small loadings on the general ability factor.

Most of the studies on test bias have involved comparisons of blacks and whites, although a number of studies involve Hispanics. I summarize here only those studies involving blacks and whites.

Test Validity

A test's predictive validity coefficient (i.e., its correlation with some criterion performance) is the most important consideration for the practical use of tests. A test with the same validity in two groups can be used with equal effectiveness in predicting the performance of individuals from each group. (The same or separate regression equations may be required for unbiased prediction, but that is a separate issue.)

The overwhelming bulk of the evidence from dozens of studies is that validity coefficients do not differ significantly between blacks and whites. In fact, other reviewers of this entire research literature have concluded that "differential validity is a nonexistent phenomenon." This conclusion applies to IQ tests for predicting scholastic performance from elementary school through high school; to college entrance tests for predicting grade-point average; to employment selection tests for predicting success in a variety of skilled, white-collar, and professional and managerial jobs; and to armed forces tests (e.g., Armed Forces Classification Test, General Classification Test) for predicting grades and successful completion of various vocational training programs.

The results of extensive test validation studies on white and black samples warrant the conclusion that today's most widely used standardized tests are just as effective for blacks as for whites in all of the usual applications of tests.

Homogeneity of Regression

Criterion performance (Y) is predicted from test scores (X) by means of a linear regression equation $\hat{Y} = a + bX$, where a is the intercept and b is the slope (which is equal to the validity coefficient when X and Y are both expressed as standardized measurements).

An important question is whether one and the same regression equation (derived from either racial group or from the combined groups) can predict the criterion with equal accuracy for members of either racial group. There are scores of studies of this question for college and employment selection tests used with blacks and whites. If the white and black regression equations do not differ in intercept and slope, the test scores can be said to have the same predictive meaning for persons regardless of whether they are black or white.

When prediction is based on a regression equation that is derived on an all-white or predominantly white sample, the results of scores of studies show, virtually without exception, one of two outcomes: (1) Usually prediction is equally accurate for blacks and whites, which means that the regressions are the same for both groups; or (2) the criterion is *over*predicted for blacks; that is, blacks do not perform as well on the criterion as their test scores predict. This is shown in Figure 3. (This finding, of course, is the opposite of the popular belief that test scores would tend to *under*estimate the criterion performance of blacks.) This predictive bias would *favor* blacks in any color-blind selection procedure. Practically all findings of predictive bias are of this type, which is called *intercept bias*, because the intercepts, but not the slopes, of the white and black regressions differ. In perhaps half of all cases of intercept bias, the bias is elminated by using "estimated true scores" instead of obtained scores. This minimizes the effect of random error of measurement, which (again, contrary to popular belief) favors the lower scoring group in any selection procedure. Improving the reliability of the test reduces the intercept bias. Increasing the *validity* of the test in both groups also reduces intercept bias. Intercept bias is a result of the test's not predicting enough of the criterion variance (in either group) to account for all of the average group difference on the criterion. Intercept bias is invariably found in those situations where the test validity is only moderate (though equal for blacks and whites) and the mean difference between groups on the criterion is as large as or almost as large as the groups' mean difference in test scores. Therefore, a test with only moderate validity cannot predict as great a difference between blacks and whites on the criterion as it should. It comes as a surprise to most people to learn that in those cases where

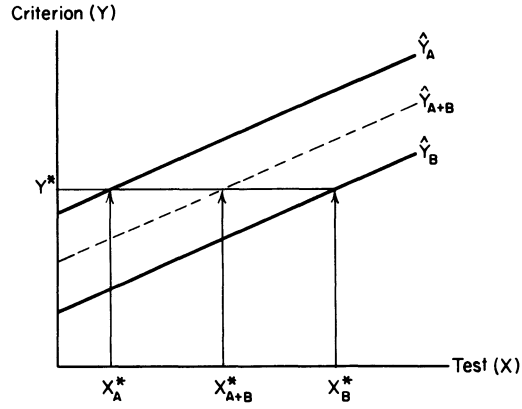

FIGURE 3. An example of the most common type of predictive bias: *intercept* bias. The major and minor groups (A and B, respectively) actually have significantly different regression lines $\hat{Y}_A$ and $\hat{Y}_B$; they differ in intercepts but not in slope. Thus, equally accurate predictions of Y can be made for individuals from either group, provided the prediction is based on the regression for the particular individual's group. If a common regression line ($\hat{Y}_{A+B}$) is used for all individuals, the criterion performance Y of individuals in Group A (the higher scoring group on the test) will be *under*predicted, and the performance of individuals in Group B (the lower scoring group) will be *over*predicted; that is, individuals in Group B will, on average, perform less well on the criterion than is predicted from the common regression line ($\hat{Y}_{A+B}$). The simplest remedy for intercept bias is to base prediction on each group's own regression line.

predictive bias is found, the bias invariably *favors* (i.e., *over*estimates) blacks. I have not come across a bona fide example of the opposite finding (Cleary, Humphreys, Kendrick, & Westman, 1975; Linn, 1973).

There are two mathematically equivalent ways to get around intercept bias: (1) Use separate regression equations for blacks and whites, or (2) enter race as a quantified variable (e.g., 0 and 1) into the regression equation. Either method yields equally accurate prediction of the criterion for blacks and whites. In the vast majority of cases, however, the intercept bias is so small (though statistically significant) as to be of no practical consequence, and many would advocate allowing the advantage of the small bias to the less favored group.

Raw Scores and Age

During the developmental period, raw scores on IQ tests show the same correlation with chronological age and the same form of growth curves for blacks as for whites.

Kinship Correlations

The correlations between twins and between full siblings are essentially the same for blacks and whites in those studies that are free of artifacts such as group differences in ceiling or floor effects, restricted range of talent, or test reliability, which can spuriously make kinship correlations unequal.

EMPIRICAL EVIDENCE ON INTERNAL INDICATORS OF BIAS

Reliability

Studies of the internal consistency reliability coefficients of standard tests of mental ability show no significant differences between whites and blacks.

Factor Analysis

When the intercorrelations among a variety of tests, such as the 11 subscales of the Wechsler Intelligence Test, the Primary Mental Abilities Tests, the General Aptitude Test Battery, and other diverse tests, are factor-analyzed separately in white and black samples, the same factors are identified in both groups. Moreover, there is usually very high "congruence" (correlation between factor loadings) between the factors in the black and white groups. If the tests measured something different in the two groups, it would be unlikely that the same factor structures and high congruence between factors would emerge from factor analysis of the tests in the two populations.

Spearman's Hypothesis

Charles Spearman originally suggested, in 1927, that the varying magnitudes of the mean differences between whites and blacks in standardized scores on a variety of mental tests were directly related to the size of the tests' loadings on *g*, the general factor common to all complex tests of mental ability. Several independent large-scale studies involving factor analysis and the extraction of a g factor from a number of diverse tests given to white and black samples show significant correlations between tests' *g* loadings and the mean white–black difference (expressed in standard score units) on the tests, thus substantiating Spearman's hypothesis. The average white–

black difference on diverse mental tests is interpreted as essentially a difference in Spearman's *g*, rather than as a difference in the more specific factors peculiar to any particular content, knowledge, acquired skills, or type of test.

Further support for Spearman's hypothesis is the finding that the average white–black difference in backward digit span (BDS) is about twice the white–black difference in forward digit span (FDS). BDS, being a cognitively more complex task than FDS, is more highly *g*-loaded (and so more highly correlated with IQ) than FDS. There is no plausible *cultural* explanation for this phenomenon (Jensen & Figueroa, 1975).

Because *g* is related to the cognitive complexity of a task, it might be predicted, in accordance with the Spearman hypothesis (that the white–black difference on tests is mainly a difference in *g*) that blacks would perform less well (relative to whites and Asians) on multiple-choice test items than on true-false items, which are less complex, having fewer alternatives to choose among. This prediction has been borne out in two studies (Longstreth, 1978).

Item × Group Interaction

This method detects a group difference in the relative difficulty of the items, determined either by analysis of the variance of the item matrix in the two groups or by correlation. The latter is more direct and easier to explain. If we determine the difficulty (percentage passing, labeled *p*) of each item of the test within each of the two groups in question, we can then calculate the correlation between the *n* pairs of *p* values (where *n* is the number of items in the test). If all the items have nearly the same rank order of difficulty in each group, the correlation between the item *p* values will approach 1.00.

The difficulty of an item is determined by a number of factors: the familiarity or rarity of its informational or cultural content, its conceptual complixity, the number of mental manipulations it requires, and so on. If the test is composed of a variety of item contents and item types, and if some items are culturally more familiar to one group than to another because of differential opportunity to acquire the different bits of information contained in different items, then we should expect the diverse items of a test to have different relative difficulties for one group and for another, if the groups' cultural backgrounds differ with respect to the informational content of the items. This, in fact, has been demonstrated. Some words in vocabulary tests have very different rank orders of difficulty for children

in England from those for children in America; some words that are common (hence easy) in England are comparatively rare (hence difficult) in America, and vice versa. This lowers the correlation of item difficulties (p values) across the two groups. If the informational demands of the various items are highly diverse, as is usually the case in tests of general ability, such as the Stanford–Binet and Wechsler scales, it would seem highly unlikely that cultural differences between groups should have a *uniform* effect on the difficulty of every item. A cultural difference would show up as differences in the rank order of item difficulties in the culturally different groups. Thus, the correlation between the rank orders of item difficulties across groups should be a sensitive index of cultural bias.

This method has been applied to a number of tests in large samples of whites and blacks. The general outcome is that the order of item difficulty is highly similar for blacks and whites and is seldom less similar than the similarity between two random halves of either the white or the black sample or between males and females of the same race. The cross-racial correlation of item difficulties determined in large samples of whites and blacks for a number of widely used standardized tests of intelligence or general ability are as follows: Stanford–Binet (.98), Wechsler Intelligence Scale for Children (.96), Peabody Picture Vocabulary Test (.98), Raven's Progressive Matrices (.98), the Wonderlic Personnel Test (.95), and the Comprehensive Tests of Basic Skills (.94). The black–white correlation of item difficulties is very much lower in tests that were intentionally designed to be culturally biased, such as the correlation of .52 found for the Black Intelligence Test (a test of knowledge of black ghetto slang terms). Because of the extremely high correlations between item difficulties for all of the standard tests that have been subjected to this method of analysis, it seems safe to conclude that the factors contributing to the relative difficulties of items in the white population are the same in the black population. That *different* factors in the two groups would produce virtually the same rank order of item difficulties in both groups would seem miraculous.

Age, Ability, and Race

It is informative to compare three types of correlations obtained within black and white populations on each of the items in a test: (1) correlation of the item with age (younger versus older children); (2) correlation of the item with ability in children of the same age as determined by total score on the test; and (3) correlation of the item with race (white versus black). We then obtain the correlations

among 1, 2, and 3 on all items. This was done for the Wechsler Intelligence Scale for Children, the Peabody Picture Vocabulary Test, and Raven's Progressive Matrices, with essentially the same results in each case: (a) The items that correlate the most with age in the black group are the same ones that correlate the most with age in the white group; (b) in both groups, the items that correlate the most with age are the same ones that correlate the most with ability; and (c) the items that correlate the most with age and ability *within* each group are the same ones that correlate the most with race. In short, the most discriminating items in terms of age and ability are the same items *within* each group, and they are also the same items that discriminate the most *between* the black and white groups. It seems highly implausible that the racial discriminability of the items, if it was due to cultural factors, would so closely mimic the item's discriminabilities with respect to age (which reflects degree of mental maturity) and ability level (with age constant) *within* each racial group.

Sociologists Gordon and Rudert (1979) have commented on these findings as follows:

> The absence of race-by-item interaction in all of these studies places severe constraints on models of the test score difference between races that rely on differential access to information. In order to account for the mean difference, such models must posit that information of a given difficulty among whites diffuses across the racial boundary to blacks in a solid front at all times and places, with no items leading or lagging behind the rest. Surely, this requirement ought to strike members of a discipline that entertains hypotheses of idiosyncratic cultural lag and complex models of idiosyncratic cultural lag and complex models of cultural diffusion (e.g., "two-step flow of communication") as unlikely. But this is not the only constraint. Items of information must also pass over the racial boundary at all times and places in order of their level of difficulty among whites, which means that they must diffuse across race in exactly the same order in which they diffuse across age boundaries, from older to younger, among both whites and blacks. These requirements imply that diffusion across race also mimics exactly the diffusion of information from brighter to slower youngsters of the same age within each race. Even if one postulates a vague but broad kind of "experience" that behaves in exactly this manner, it should be evident that would represent but a thinly disguised tautology for mental functions that IQ tests are designed to measure. (pp. 179–180)

Verbal versus Nonverbal Tests

Because verbal tests, which, of course, depend on specific language, would seem to afford more scope for cultural influences than

nonverbal tests, it has been commonly believed that blacks would score lower on verbal than on nonverbal tests.

A review of the entire literature comparing whites and blacks on verbal and nonverbal tests reveals that the opposite is true: Blacks score slightly better on verbal than on nonverbal tests. However, when verbal and nonverbal items are all perfectly matched for difficulty in white samples, blacks show no significant difference on the verbal and nonverbal tests. Hispanics and Asians, on the other hand, score lower on verbal than on nonverbal tests.

The finding that blacks do better on tests that are judged to be more culture-loaded than on tests judged to be less culture-loaded can be explained by the fact that the most culture-loaded tests are less abstract and depend more on memory and recall of past-acquired information, whereas the least culture-loaded tests are often more abstract and depend more on reasoning and problem solving. Memory is less *g*-loaded than reasoning, and so, in accord with Spearman's hypothesis, the white–black difference is smaller on tests that are more dependent on memory than on reasoning.

DEVELOPMENT TESTS

A number of tests devised for the early childhood years are especially revealing of both the quantitative and the qualitative features of cognitive development—such as Piaget's specially contrived tasks and procedures for determining the different ages at which children acquire certain basic concepts, such as the conservation of volume (i.e., the amount of liquid is not altered by the shape of its container) and the horizontality of liquid (the surface of a liquid remains horizontal when its container is tilted). Black children lag one to two years behind white and Asian children in the ages at which they demonstrate these and other similar concepts in the Piagetian tests, which are notable for their dependence only on things that are universally available to experience.

Another revealing developmental task is copying simple geometric figures of increasing complexity (e.g., circle, cross, square, triangle, diamond, cylinder, cube). Different kinds of copying errors are typical of different ages; black children lag almost two years behind white and Asian children in their ability to copy figures of a given level of complexity, and the nature of their copying errors is indistinguishable from that of white children about two years younger.

White children lag about six months behind Asians in both the Piagetian tests and the figure-copying tests.

Free drawings, too, can be graded for mental maturity, which is systematically reflected in such features as the location of the horizon line and the use of perspective. Here, too, black children lag behind the white.

A similar developmental lag is seen also in the choice of error distractors in the multiple-choice alternatives on Raven's Progressive Matrices, a nonverbal reasoning test. The most typical errors made on the Raven test systematically change with the age of children taking the test, and the errors made by black children of a given age are typical of the errors made by white children who are about two years younger.

In a "test" involving only preferences of the stimulus dimensions selected for matching figures on the basis of color, shape, size, and number, 5- to 6-year-old black children show stimulus-matching preferences typical of younger white children.

In summary, in a variety of developmental tasks, the performance of black children at a given age is quantitatively and qualitatively indistinguishable from that of white and Asian children who are one to two years younger. The consistency of this lag in capability, as well as the fact that the typical qualitative features of blacks' performance at a given age do not differ in any way from the features displayed by younger white children, suggests that this is a developmental rather than a cultural effect.

PROCEDURAL AND SITUATIONAL SOURCES OF BIAS

A number of situational variables external to the tests themselves, which have been hypothesized to influence test performance, were examined as possible sources of bias in the testing of different racial and social class groups. The evidence is wholly negative for every such variable on which empirical studies are reported in the literature. That is to say, no variables in the test situation have been identified that contribute significantly to the observed average test-score differences between social classes and racial groups.

Practice effects in general are small, amounting to a gain of about 5 IQ points between the first and second test, and becoming much less thereafter. Special coaching on test-taking skills may add another 4–5 IQ points (over the practice effect) on subsequent tests if these are highly similar to the test on which subjects were coached. However,

neither practice effects nor coaching interacts significantly with race or social class. These findings suggest that experience with standard tests is approximately equal across different racial and social class groups. None of the observed racial or social class differences in test scores is attributable to differences in amount of experience with tests *per se*.

A review of 30 studies addressed to the effect of the race of the tester on test scores reveals that this is preponderantly nonsignificant and negligible. The evidence conclusively contradicts the hypothesis that subjects of either race perform better when tested by a person of the same race than when tested by a person of a different one. In brief, the existence of a race of examiner × race of subject interaction is not substantiated.

The language style or dialect of the examiner has no effect on the IQ performance of black children or adults, who do not score higher on verbal tests translated and administered in black ghetto dialect than on those in standard English. On the other hand, all major *bilingual* populations in the United States score slightly but significantly lower on verbal tests (in standard English) than on nonverbal tests, a finding suggesting that a specific language factor is involved in their lower scores on verbal tests.

The teacher's or tester's expectation concerning the child's level of ability has no demonstrable effect on the child's performance on IQ tests. I have found no bona fide study in the literature that shows a significant expectancy (or "Pygmalion") effect for IQ.

Significant but small "halo effects" on the *scoring* of subjectively scored tests (e.g., some of the verbal scales of the Wechsler) have been found in some studies, but these halo effects have not been found to interact with either the race of the scorer or the race of the subject.

Speeded versus unspeeded tests do not interact with race or social class, and the evidence contradicts the notion that speed or time pressure in the test situation contributes anything to the average test-score differences between racial groups or social classes. The same conclusion is supported by evidence concerning the effects of varying the conditions of testing with respect to instructions, examiner attitudes, incentives, and rewards.

Test anxiety has not been found to have differential effects on the test performances of blacks and whites. Studies of the effects of achievement motivation and self-esteem on test performance also show largely negative results in this respect.

In summary, as yet no factors in the testing procedure itself have been identified as sources of bias in the test performances of different racial groups and social classes.

OVERVIEW

Good tests of abilities surely do not measure human worth in any absolute sense, but they do provide indices that are correlated with certain types of performance generally deemed important for achieving responsible and productive roles in our present-day society.

Most current standardized tests of mental ability yield unbiased measures for all native-born English-speaking segments of American society today, regardless of their sex or their racial and social class background. The observed mean differences in test scores between various groups are generally not an artifact of the tests themselves but are attributable to factors that are causally independent of the tests. The constructors, publishers, and users of tests need to be concerned only about the psychometric soundness of these instruments and must apply appropriate objective methods for detecting any possible biases in test scores for the groups in which they are used. Beyond that responsibility, the constructors, publishers, and users of tests are under no obligation to explain the *causes* of the statistical differences in test scores between various subpopulations. They can remain agnostic on that issue. Discovery of the causes of the observed racial and social-class differences in abilities is a complex task calling for the collaboration of several specialized fields in the biological and behavioral sciences, in addition to psychometrics.

Whatever may be the causes of group differences that remain after test bias is eliminated, the practical applications of sound psychometrics can help to reinforce the democratic ideal of treating every person according to the person's *individual* characteristics, rather than according to his or her sex, race, social class, religion, or national origin.

SECOND THOUGHTS ON *BIAS IN MENTAL TESTING*

More than 100 reviews, critiques, and commentaries have been addressed to my *Bias in Mental Testing* since its publication in Jan-

uary 1980. (A good sampling of 27 critiques, including my replies to them, is to be found in the "Open Peer Commentary" in *Brain and Behavioral Sciences,* 1980, 3, 325–371.) It is of considerable interest that not a single one has challenged the book's main conclusions, as summarized in the preceding section. This seemed to me remarkable, considering that these conclusions go directly counter to the prevailing popular notions about test bias. We had all been brought up with the conviction that mental ability tests of nearly every type are culturally biased against all racial and ethnic minorities and the poor and are slanted in favor of the white middle class. The contradiction of this belief by massive empirical evidence pertinent to a variety of criteria for directly testing the cultural bias hypothesis has revealed a degree of consensus about the main conclusions that seems unusual in the social sciences: The observed differences in score distributions on the most widely used standardized tests between native-born, English-speaking racial groups in the United States are not the result of artifacts or shortcomings of the tests themselves; they represent real differences—*phenotypic* differences, certainly—between groups in the abilities, aptitudes, or achievements measured by the tests. I have not found any critic who, after reading *Bias in Mental Testing,* has seriously questioned this conclusion, in the sense of presenting any contrary evidence or of faulting the essential methodology for detecting test bias. This is not to suggest that there has been a dearth of criticism, but criticisms have been directed only at a number of side issues, unessential to the cultural bias hypothesis, and to technical issues in factor analysis and statistics that are not critical to the main argument. But no large and complex work is unassailable in this respect.

Of all the criticisms that have come to my attention so far, are there any that would cause important conceptual shifts in my thinking about the main issues? Yes, there are several important points that I am now persuaded should be handled somewhat differently if I were to prepare a revised edition of *Bias.*

Generalizability of Predictive Validity

The belief that the predictive validity of a job selection test is highly specific to the precise *job,* the unique *situation* in which the workers must perform, and the particular *population* employed has been so long entrenched in our thinking as to deserve a special name. I shall call it the *specificity doctrine.* This doctrine has been incorporated as a key feature of the federal "Uniform Guidelines on

Employee Selection Procedures" (Equal Employment Opportunity Commission, 1978), which requires that where tests show "adverse impact" on minority hiring or promotion because of average majority–minority differences in test scores, the predictive validity of the tests must be demonstrated for each and every job in which test scores enter into employee selection. In *Bias,* I had given rather uncritical acceptance to this doctrine, at least as it regards job specificity, but I have since learned of the extremely important research of John E. Hunter and Frank L. Schmidt and their co-workers, cogently demonstrating that the specificity doctrine is false (e.g., Schmidt & Hunter, 1977). This doctrine gained currency because of failure to recognize certain statistical and psychometric artifacts, mainly the large sampling error in the many typical small-sample validity studies. When this error-based variability in the validity coefficients for a given test, as used to predict performance in a variety of jobs in different situations in different populations, is properly taken into account, the specificity doctrine is proved false. Most standard aptitude tests, in fact, have the same true validity across many jobs within broad categories of situations and subpopulations. Schmidt and Hunter (1981) based their unequivocal conclusions on unusually massive evidence of test validities for numerous jobs. They stated, "The theory of job specific test validity is false. Any cognitive ability test is valid for any job. There is no empirical basis for requiring separate validity studies for each job" (p. 1133).

In *Bias,* I also gave too much weight to the distinction between test validity for predicting success in job training and later actual performance on the job. But this turns out to be just another facet of the fallacious specificity doctrine. Again, a statistically proper analysis of the issue led Schmidt and Hunter (1981) to this conclusion:

> Any cognitive test valid for predicting performance in training programs is also valid for predicting later performance on the job . . . when employers select people who will do well in training programs, they are also selecting people who will do well later on the job. (p. 1133)

Differential Validity for Majority and Minority Groups

Although the vast majority of studies of the predictive validity of college entrance tests and personnel selection tests shows nonsignificantly different validity coefficients, regressions, and standard errors of estimate in white and black and Hispanic samples, there are occasionally statistically significant differences between the groups in these parameters. I now believe I did not go far enough in putting

these relatively few deviant findings in the proper perspective, statistically. To do so becomes possible, of course, only when a large number of studies is available. Then, as Hunter and Schmidt (e.g., 1978) have pointed out repeatedly in recent years, we are able to estimate the means and standard deviations of the various validity parameters over numerous studies in the majority and the minority, and by taking proper account of the several statistical artifacts that contribute to the between-studies variability of these parameters, we can better evaluate the most deviant studies. Such meta-analysis of the results of numerous studies supports an even stronger conclusion of the general absence of bias in the testing of minorities than I had indicated in my book. When subjected to meta-analysis, the few deviant studies require no special psychological or cultural explanations; they can be interpreted as the tail ends of the between-studies variation that is statistically assured by sampling error and differences in criterion reliability, test reliability, range restriction, criterion contamination, and factor structure of the tests. Taking these sources of variability into account in the meta-analysis of validity studies largely undermines the supposed importance of such moderator variables as ethnic group, social class, sex, and geographic locality. I hope that someone will undertake a thorough meta-analysis of the empirical studies of test bias, along the lines suggested by Hunter and Schmidt (e.g., Schmidt, Hunter, Pearlman, & Shane, 1979). Their own applications of meta-analysis to bias in predictive validities has led to very strong conclusions, which they have clearly spelled out in the present volume. When applied to other types of test bias studies, such as groups-by-items interaction, I suspect it will yield equally clarifying results. These potentially more definitive meta-analytic conclusions are latent, although not objectively explicit, in my own summaries of the evidence in *Bias*, which in some ways probably *understated* the case that most standard tests are culturally unbiased for American-born racial and ethnic minorities.

Bilingualism and Verbal Ability

A recent article by sociologist Robert A. Gordon (1980), which appeared after *Bias*, is one of the most perceptive contributions I have read in the test bias literature. One point in Gordon's article (pp. 177–180) especially gave me pause. Until I read it, I had more or less taken for granted what seemed the commonsense notion that *verbal* tests are biased, or at least highly suspect of that possibility, for any

bilingual person, particularly if the verbal test is in the person's second language. But Gordon pointed out that bilingualism and low verbal ability (relative to other abilities), independent of any specific language, may covary across certain subpopulations merely by happenstance, and that not all of the relative verbal-ability deficit is causally related to bilingualism *per se*. The educational disadvantage of bilingualism may be largely the result of lower verbal aptitude *per se* than of a bilingual background. Admittedly, it is psychometrically problematic to assess verbal ability (independently of general intelligence) in groups with varied language backgrounds. But Gordon has made it clear to me, at least, that we cannot uncritically assume that bilingual groups will necessarily perform below par on verbal tests, or that, if they do, the cause is necessarily their bilingualism. Gordon noted some bilingual groups that perform better, on the average, on verbal tests *in their second language* than on nonverbal reasoning tests. Samples from certain ethnic groups that are entirely monolingual, with no exposure to a second language, nevertheless show considerable differences between levels of verbal and nonverbal test performance. Gordon hypothesized that acquisition of English would proceed most rapidly among immigrant groups natively high in verbal ability, which would lead eventually to a confounding between low verbal ability and bilingual handicap. He noted, for example, that verbal IQ had no relation to degree of bilingualism among American Jews, once the children were several years in public school. Such findings would seem to call for a more thorough and critical assessment of the meaning of lower verbal test scores in today's predominant bilingual groups in America.

Interpretation of Groups × Item Interaction as a Detector of Cultural Bias

The statistical interaction of group × item in the analysis of variance (ANOVA) of the total matrix of groups, subjects, and items has been one of the most frequently used means of assessing item bias in tests. The method is very closely related to another method of assessing item bias, the correlation (Pearson r) between the item p values (percentage of each group passing each item) of the two population groups in question. A perfect correlation between the groups' p values is the same as a group × item interaction of zero, and there is a perfect inverse relationship between the size of the correlation between groups' p values and the size of the group × item interaction term in the complete ANOVA of the group × item × subject

matrix. The advantage of the correlation method is that it yields, in the correlation coefficient, a direct indication of the degree of similarity (with respect to both rank order and interval properties) of the item *p* values in the two groups, for example, whites and blacks. The advantage of the ANOVA group × item interaction method is that it provides a statistical test of the significance of the group difference in the relative difficulties of the items.

Applications of both methods to test data on whites and blacks have generally shown very high correlations ($r > .95$) between the groups' *p* values. The group × item interaction is usually very small relative to other sources of variance (usually less than 1% or 2% of the total variance), but it is often statistically significant when the sample size is large ($N > 200$). It has also been observed that if the comparison groups (usually blacks and whites) are composed of subjects who are specially selected on the basis of total scores so as to create black and white groups that are perfectly matched in overall ability, the correlation between the matched groups' *p* values is even higher than the correlation for unmatched groups, and (in the ANOVA of the matched groups) the group × item interaction is appreciably reduced, usually to nonsignificance.

Some critics have interpreted this finding as an indication that the black and white groups that are matched on overall ability (e.g., total test score) show a smaller group × item interaction because they have developed in culturally more similar backgrounds than the unmatched samples. However, this is not necessarily so. There is no need to hypothesize cultural differences to explain the observed effects—at least, no cultural factors that would cause significant group × item interaction. The observed group × item interaction, in virtually all cases that we have examined, turns out to be an artifact of the method of scaling item difficulty. Essentially, it is a result of the nonlinearity of the item-characteristic curve. As I failed to explain this artifact adequately in my treatment of the group × item method in *Bias in Mental Testing*, I will attempt to do so here.

A hypothetical simplest case is shown in the item-characteristic curves (ICC) of Figure 4. Assume that the ICC of each item, *i* and *j*, is *identical* for the two populations, A and B. The ICC represents the percentage of the population passing a given item as a function of the overall ability (*X*) measured by the test as a whole. If an item's ICC is identical for the two populations, it means that the item is an unbiased measure of the same ability in both groups; that is, the item is related to ability in the same way for members of both groups. When two groups' ICCs are the same, individuals of a given level of

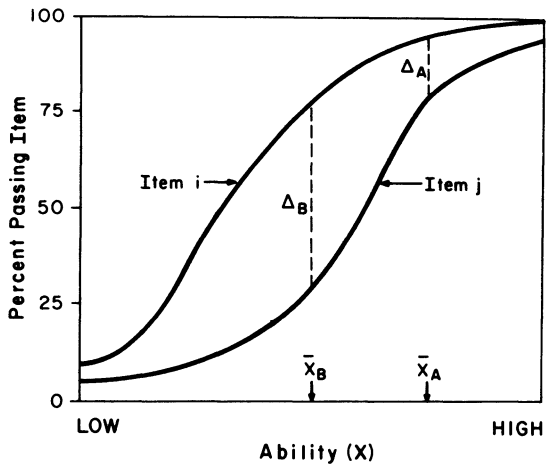

FIGURE 4. Hypothetical item-characteristic curves (ICC) for items *i* and *j*, illustrating the typically nonlinear relationship between probability of a correct response to the test item and the ability level of persons attempting the item.

ability X will have the same probability of passing a given item, regardless of their group membership. This is one definition of an unbiased item. Therefore, in our simple example in Figure 4, both items, *i* and *j*, are unbiased items. Yet, they can be seen to show a significant group $\times$ item interaction. But this interaction is an artifact of the nonlinearity of the ICCs. The ICC is typically a logistic or S-shaped curve, as shown in Figure 4. If the means, $\bar{X}_A$ and $\bar{X}_B$, of two groups, A and B, are located at different points on the ability scale, and if any two items, *i* and *j*, have different ICCs (as is always true for items that differ in difficulty), then, the difference Δ_A between the percentage passing items *i* and *j* in group A will differ from the difference Δ_B between the percentage passing items *i* and *j* in Group B. This, of course, is what is meant by a group $\times$ item interaction; that is, Δ_B is significantly greater than Δ_A. If the ordinate (in Figure 4) were scaled in such a way as to make the two ICCs perfectly linear and parallel to one another, there would be no interaction. There could be no objection to changing the scale on the ordinate, as p (percentage passing) is just an arbitrary index of item difficulty. It can be seen from Figure 4 that matching the groups on ability so that $\bar{X}_A = \bar{X}_B$ will result in exactly the same Δ for both groups (i.e., no group $\times$ item interaction).

The practical implication of this demonstration for all data that now exist regarding group $\times$ item interaction is that the small but

significant observed group × item interactions would virtually be reduced to nonsignificance if the artifact due to ICC nonlinearity were taken into account. It is likely that the correct conclusion is that in most widely used standard tests administered to any American-born English-speaking populations, regardless of race or ethnic background, group × item interaction is either trivially small or a nonexistent phenomenon.

This conclusion, however, does not seem to me to be a trivial one, as Jane Mercer claims. The fact that item-characteristic curves on a test like the Scholastic Aptitude Test (SAT) are the same (or nonsignificantly different) for majority and minority groups in the United States runs as strongly counter to the cultural-bias hypothesis as any finding revealed by research. To argue otherwise depends on the implausible hypothesis that the cultural difference between, say, blacks and whites affects every item equally, and that the cultural disadvantage diffuses across all items in a uniform way that perfectly mimics the effects on item difficulty of differences in ability level *within* either racial group, as well as differences in chronological age *within* either racial group. A much more plausible hypothesis is that either (1) the cultural differences between the racial groups are so small as not to be reflected in the item statistics, or (2) the items composing most present-day standardized tests have been selected in such a way as not to reflect whatever differences in cultural backgrounds may exist between blacks and whites. If test items were typically as hypersensitive to cultural differences (real or supposed) as some test critics would have us believe, it is hard to imagine how such a variety of items as is found in most tests would be so *equally* sensitive as to show Pearsonian correlations between blacks and white item difficulties (p values) in the upper .90s. And even these very high correlations, as explained previously, are attenuated by the nonlinearity of the ICCs. The total evidence on item bias, in numerous well-known tests, gives no indication of a distinctive black culture in the United States.

Methods of Factor Analysis

Because all the intercorrelations among ability tests, when obtained in a large representative sample of the general population, are *positive,* indicating the presence of a general factor, I believe that it is psychologically and theoretically wrong to apply any method of factor analysis in the abilities domain that does not permit estimation of the general factor. Methods of factor analysis involving orthogonal

rotation of the factor axes, which submerges the general factor, may make as much sense mathematically as any other methods of factor analysis, but they make much less sense psychologically. They ignore the most salient feature of the correlation matrix for ability tests: positive manifold.

In *Bias,* I considered various methods of extracting *g* and the group factors. This is not the appropriate place to go into all of the technical details on which a comparison of the various methods must depend. But now, I would emphasize, more than I did in *Bias,* that in my empirical experience, the *g* factor is remarkably robust across different methods of extraction on the same set of data, and it is also remarkably robust across different populations (e.g., male and female, and black and white). The robustness, or invariance, of *g* pertains more to the *relative* magnitudes and rank order of the individual tests' *g* loadings than to the absolute amount of variance accounted for by the *g* factor. The first principal component accounts for the most variance; the first principal factor of a common factor analysis accounts for slightly less variance; and a hierarchical or second-order *g*, derived from the intercorrelations among the obliquely rotated first-order factors, accounts for still less of the total variance. But the rank orders of the *g* loadings are highly similar, with congruence coefficients generally above .95, among all three methods of *g* extraction. This has been found in more than two dozen test batteries that I have analyzed, each test by all three methods. This outcome, however, is not a mathematical necessity. Theoretically, collections of tests could be formed that would yield considerably different *g* factors by the different methods. This would occur when a particular type of ability test is greatly overrepresented in the battery in relation to tests of other abilities. The best insurance against this possible distortion of *g* is a hierarchical analysis, with *g* extracted as a second-order factor.

Rotation of factor axes is often needed for a clear-cut interpretation of the factors beyond the first (which is usually interpreted as *g*). In *Bias* (p. 257), I suggested taking out the first principal factor and then orthogonally rotating the remaining factors (plus one additional factor), using Kaiser's varimax criterion for approximating simple structure. This suggested method is inadequate and will be deleted in subsequent printings and editions of *Bias.* A mathematically more defensible method, and one that I find empirically yields much clearer results, had already been devised (Schmid & Leiman, 1957; and Wherry, 1959, using a different computational routine leading to the same results). The Schmid–Leiman method is hierarchical; it

extracts first-order oblique factors, and from the intercorrelations among these, it extracts a second-order (or other higher order) *g* factor; and then the first-order oblique factors are "orthogonalized"; that is, with the *g* removed to a higher level, the first-order factors are left uncorrelated (i.e., orthogonal). The Schmid–Leiman transformation, as it is known, now seems to me to result in the clearest, theoretically most defensible, factor-analytic results in the ability domain. Like all hierarchical solutions, the Schmid–Leiman transformation is probably more sensitive to statistical sampling error than are principal components and common factor analysis, and so its wise use depends on reasonably large samples. The Schmid–Leiman transformation warrants greater recognition and use in the factor analysis of ability tests. In the study of test bias, it seems an optimal method for comparing the factor structures of a battery of tests in two or more subpopulations, provided the sample sizes are quite large ($N > 200$).

Genotypes and Phenotypes

I stated in the preface of *Bias,* and again in my final chapter, that the study of test bias is *not* the study of the heredity–environment question, and that the findings on bias cannot explain the cause of group differences, except to rule out test bias itself as a possibile cause. I emphasized that all that tests can measure directly are phenotypes: *All test scores are phenotypes.* The chief aim of the study of test bias is to determine whether the measurements of phenotypic differences are biased. That is, are they an artifact of the measurement technique *per se,* or do they reflect real phenotypic differences in a broader sense, with implications beyond the test scores themsleves? My analysis of the massive evidence on this issue led me to conclude in *Bias,* "The observed mean differences in test scores between various [racial and social class] groups are generally not an artifact of the tests themselves, but are attributable to factors that are causally independent of the tests" (p. 740).

Despite my clearly stated position regarding the study of test bias in relation to the heredity–environment question, a number of critics and reviewers, in this volume and elsewhere (e.g., "Open Peer Commentary," 1980), have insisted on discussing heredity–environment in the context of test bias. It makes me think that perhaps I have not stated my thoughts on this matter strongly and fully enough in *Bias.* I will try to do so here.

Misunderstandings on this issue fall into two main categories: (*a*) Nonbiased test scores mean genetic differences, and (b) if group differences are not proved to be genetic, they are not really important. Both propositions are clearly false, but we must examine them more closely to see why.

a. First, let us look at the belief that if a test has been shown to be unbiased, any group difference in test scores must be due to genetic factors. The primary fallacy here is the implicit assumption that a test's bias (or absence of bias) applies to *every* criterion that the test might conceivably be used to predict. A test score (X) is said to be biased with respect to two (or more) groups if it either overpredicts or underpredicts a criterion measurement (Y) for one group when prediction is based on the common regression of Y on X in the two (or more) groups. But there is nothing in the logic of psychometrics or statistical regression theory that dictates that a test that is biased (or unbiased) with respect to a particular criterion is necessarily biased (or unbiased) with repsect to some other criterion. Whether a test is or is not biased with respect to some other criterion is a purely empirical question. It is merely an empirical fact, not a logical or mathematical necessity, that a test that is found to be an unbiased predictor of one criterion is also generally found to be an unbiased predictor of many other criteria—usually somewhat similar criteria in terms of their factorial composition of requisite abilities. But the genotype is conceptually quite different from the criteria that test scores are ordinarily used to predict—such criteria as school and college grades, success in job-training programs, and job performance. Some critics have been overly defensive about the general finding of nonbias in so many standard tests for blacks and whites with respect to the criterion validity and other external correlates of the test scores, which they have apparently viewed as presumptive evidence that the scores are probably also unbiased estimators of intelligence genotypes in different racial groups. This may seem a plausible inference; it is certainly not a logical inference. The issue is an empirical one. I have not found any compelling evidence marshaled with respect to it. As I have explained in greater detail elsewhere (Jensen, 1981), answers to the question of the relative importance of genetic and nongenetic causes of the average differences between certain racial groups in test performance (and all the correlates of test performance) at present unfortunately lie in the limbo of mere plausibility and not in the realm of scientific verification. Without a true genetic experiment, involving cross-breeding of random samples of racial populations in every race $\times$ sex combination, as well

as the cross-fostering of the progeny, all currently available types of test results and other behavioral evidence can do no more than enhance the plausibility (or implausibility) of a genetic hypothesis about any particular racial difference. Whatever social importance one may accord to the race–genetics question regarding IQ, the problem is scientifically trivial, in the sense that the means of answering it are already fully available. The required methodology is routine in plant and animal experimental genetics. It is only because this appropriate well-developed methodology must be ruled out of bounds for social and ethical reasons that the problem taxes scientific ingenuity and may even be insoluble under these constraints.

Although it is axiomatic that test scores are measures of the phenotype only, this does not preclude the *estimation* of individuals' genotypes from test scores, given other essential information. One can see the logic of this estimation, using the simplest possible quantitative-genetic model:

$$P = G + E$$

where P is the individual's phenotypic deviation from the mean, $\overline{P}$, of all the individual phenotypic values in the population of which the individual is a member; G is the individual's genotypic deviation from the mean genetic effect in the population; and E is the individual's deviation from the mean environmental effect in the population. The (broad) *heritability*, h^2, of P in the population is defined as the squared correlation between phenotypic and genotypic values, that is, $h^2 = r^2_{PG}$. Methods of quantitative genetics, using a variety of kinship correlations, can estimate h^2. (For mental test scores, most estimates of h_2 in numerous studies fall in the range from .50 to .80). If we assume, for the sake of expository simplicity, that h^2 can be determined without sampling error, then it follows from our statistical model that we can obtain an estimate, $\hat{G}$, of an individual's genotypic value, G, given P for that individual: $G = h^2P$. The $\hat{G}$, of course, has a standard error of estimate, just as any other value estimated from a regression equation. In this case, the error of estimate for $\hat{G}$ is $\sigma_P h\sqrt{1 - h^2}$, where σ_P is the standard deviation of P in the population.

It is seen that all the parameters involved in this estimation procedure are specific to the population of which the individual is a member. Therefore, although the statistical logic of G estimation permits us to compare the $\hat{G}$ values of individuals from the same population, and to test the difference between individuals for statistical significance at some specified level of confidence, it cannot logically

justify the comparison of $\hat{G}$ values of individuals from *different* populations, even if h^2 is identical *within* each of the two populations. In other words, the logic of estimation of G from this model *within* a given population cannot be extended to the mean difference *between* two populations. Here is why: If $\overline{P}_C$ = the mean of two populations, A and B, combined and $\overline{P}_A$ and $\overline{P}_B$ are the deviations of the population means (on the phenotype) from the composite mean, $\overline{P}_C$, then the calculation of $\hat{G}_A$ or $\hat{G}_B$ from the model described above would be $\hat{G}_A = h^2\overline{P}_A$ and $G_B = h^2\overline{P}_B$. But in this case, the required h^2 is *not* the h^2 *within* each population (or *within* the combined populations), as in estimating G for individuals; what is required is the heritability of the difference *between* the two populations. But we have no way of determining h^2 *between* populations, short of a true genetic experiment involving random cross-breeding and cross-fostering of the two populations. Thus, if the means, $\overline{P}_A$ and $\overline{P}_B$, of two populations, A and B, differ on a given scale, we cannot infer whether it is because $G_A \neq G_B$, or $E_A \neq E_B$, or some weighted combination of these component differences, and this limitation is as true of measurements of height or weight or any other physical measurements as it is of mental test scores: They are all just *phenotypes*, and the logic of quantitative genetics applies equally to all metric traits. If you believe that Watusis are taller than Pygmies because of genetic factors, it is only because this belief seems plausible to you, not because there is any bona fide genetic evidence for it. We are in essentially the same position regarding racial differences in mental test scores. The mistake is to assume, in the absence of adequate evidence, that either the plausible or the opposite of the plausible is true. All that we mean by *true* in the scientific sense is that the evidence for a given conclusion is deemed adequate by the current standards of the science. By the standards of genetics, adequate evidence for a definitive conclusion regarding the race–genetics mental ability question is not at hand. In the absence of adequate evidence, the only defensible posture for a scientist is to be openly agnostic. Unfortunately, it is often more dangerous to be openly agnostic about the race–IQ–genetics question than to be loudly dogmatic on the environmentalist side.

The fact that a genetic difference between two populations cannot properly be inferred on the basis of estimates of h^2 in both populations, however, should not be misconstrued, as it so often is, to mean that the heritability of a trait *within* each of two groups has no implication whatsoever with respect to the causes of the mean difference *between* the groups. To make the explanation simple, consider the case of complete heritability ($h^2 = 1$) *within* each of two

groups for which the distributions of measurable phenotypes have different means. The fact is that $h^2 = 1$ severely constrains the possible explanations of the causes of the mean difference between the groups. It means that none of the environmental (or nongenetic) factors showing variation *within* the groups could be the cause of the group difference if the groups are, in fact, not genetically different. It would mean either (a) that the groups differ genetically or (b) that the group difference is the result of some nongenetic factor(s) not varying among individuals within either group, or both (a) and (b). To the extent that the heritability within groups increasingly exceeds zero, heritability implies some increasing constraint on the environmental explanation of a difference between the groups, the degree of constraint also being related to both the magnitude of the mean difference and the amount of overlap of the two phenotypic distributions. Within-group heritability *per se,* whatever its magnitude, of course, could never demonstrate heritability between groups. But no knowledgeable person has ever claimed that it does.

b. If a phenotypic difference between groups cannot be attributed to genetic factors, or if its cause is unknown, is it therefore unimportant? Not at all. There is no necessary connection at all between the individual or social importance of a phenotypic trait and its degree of heritability. The importance of variation on any trait or behavior must be judged in terms of its practical consequences for the individual and for society, regardless of the causes of such variation. For many years now, there has been a very broad consensus that the IQ deficit of black Americans is important—not because performance on an IQ test *per se* is important, but because of all of the "real-life" behavioral correlates of the IQ that are deemed important by society, and these correlations are largely the same for blacks as for whites. The complete disappearance of mental tests would not in the least diminish all of the educational, occupational, and economic consequences of the fact that, at this time, black Americans, on average, are about one standard deviation below the white and Asian populations in general mental ability. The immediate practical consequences of this deficit are the same, whether or not we understand its cause. What we *do* know, at present, is that mental tests are not the cause of the deficit, but merely an accurate indicator of it.

Lloyd Humphreys (1980a) has written tellingly on this point. He concluded:

> The phenotypic difference is important, not trivial. It is real, not ephemeral. It is not a spurious product of the tests and the test-taking situation

> but extends to classrooms and occupations. Today the primary obstacle to the achievement by blacks of proportional representation in higher education and in occupations is not the intelligence test or any of its derivatives. Instead, it is the lower mean level of black achievement in basic academic, intellectual skills at the end of the public school period. It is immaterial whether this mean deficit is measured by an intelligence test, by a battery of achievement tests, by grades in integrated classrooms, or by performance in job training. The deficit exists, it is much broader than a difference on tests, and there is no evidence that , even if entirely environmental in origin, it can be readily overcome. From this point of view it is immaterial whether the causes are predominantely genetic or environmental. (pp. 347–348)

COMMENTARY ON PREVIOUS CHAPTERS

From here on, I will comment on specific points that have especially attracted my attention in the other contributions to this volume, taking the chapters in alphabetical order by first author. Naturally, I have the most to say about those chapters in which I find some basis for disagreement. I see little value in noting all the points of agreement.

BERNAL

Bernal's main argument is that something he refers to as the "total testing ambience" has the effect of depressing the test performance of minority subjects. Although the meaning of *testing ambience* is not made entirely clear, it presumably involves certain attitudes and skills that are amenable to teaching or to an experimental manipulation of the test situation. It is not a novel idea, and there is a considerable empirical literature on it. The best studies I could find in the literature are reviewed in Chapter 12 ("External Sources of Bias") in *Bias in Mental Testing*. The reviewed studies have taken account of practice effects on tests, interpersonal effects (race, attitude, expectancy, and dialect of examiner), manner of giving test instructions, motivating and rewarding by the examiner, individual and group administration, timed versus untimed tests, and the effects of classroom morale and discipline on test performance. The overwhelming conclusion from all these studies is that these "ambience" variables make a nonsignificant and negligible contribution to the observed racial and social class differences in mean test scores on standardized tests. If there are published studies that would lead to a

contrary conclusion, I have not been able to find them, and Bernal has not cited them.

Bernal states, "As in his previous works, Jensen continued to use selected studies to broaden the data base that supports his basic contentions" (Chap. 5, p. 172). Actually, in *Bias,* I was not selective of the studies I cited; I tried to be as comprehensive as feasibly possible in reviewing relevant studies. If I have overlooked relevant studies, then these should be pointed out, with a clear explanation of how their results would alter my conclusions based on the studies I reviewed. In all the reviews and critiques of *Bias* since its publication two years ago, I have not seen any attempt to bring forth any evidence that I may have overlooked and that would contradict any of my main conclusions. If Bernal (and Hilliard) know of any such evidence, they have kept it a secret.

Elsewhere (Jensen, 1976), I have explained why it is logically fallacious to infer either test bias or the absence of genetic effects from the presence or absence of training effects on test performance. The demonstration of a training effect on a particular trait or skill is not at all incompatible either with nonbias in the test measuring the skill (before *or* after training) or with a high degree of genetic determination of individual or group differences. An experiment involving a group $\times$ training design does not logically permit conclusions concerning the genetic or nongenetic causes of the main effect of the group difference or their interaction with treatments, nor can such a design reflect on the culture-fairness of the measuring instrument. But this restriction of inference about bias applies only to training subjects in the ability, knowledge, or skill measured by the test itself. It should not apply to the testing ambience, which includes the instructions for taking the test and the atmosphere in which it is administered. It is important that all subjects understand the instructions and the sheer mechanics of taking the test. When these situational factors have been experimentally manipulated, however, they have generally shown small but statistically significant main effects of the experimental treatment, but they have not shown significant interactions with race or social class (see Jensen, 1980a, pp. 611–615). We shall see if Bernal's own experiment is an exception to this general finding.

But first, two other more general observations are called forth by Bernal's chapter.

Bernal refers mainly to children in test situations, for it is in this age group that lack of sophistication in test taking is most likely. But the white–black differences in test performance observed among ele-

mentary-school children are no greater, in standard score units, than the racial differences seen between much older groups that have become much more test-wise, after completing 12 years of public school, or 4 years of college, or an additional 3 or 4 years of post-graduate professional school. Yet, differences of one standard deviation or more are found between whites and blacks on the Armed-Forces Qualification Test, on college entrance exams such as the SAT, on the Graduate Record Exam (taken after college graduation), on the Law School Admission Test and the Medical College Aptitude Test (taken after prelaw and premedical college programs), and on state bar exams (taken after graduation from law school), which, according to the National Bar Association, are failed by three out of four black law school graduates—a rate two to three times that of their white counterparts. Data provided by the test publishers on these various post-high-school tests, based on nationwide test scores obtained in recent years, are summarized in Table 1 in terms of the mean difference between the white and minority groups, expressed in standard deviation units (i.e., the mean difference divided by the

TABLE 1
Mean Difference (in Standard Deviation Units) between Whites and Blacks (W–B) and Whites and Chicanos (W-C) on Various College and Postgraduate Level Tests[a]

Test	Difference in *SD* units	
	W–B	W–C
Scholastic Aptitude Test—Verbal	1.19	0.83
Scholastic Aptitude Test—Math	1.28	0.78
American College Test	1.58	1.22
National Merit Qualifying Exam.	1.11	
Graduate Record Exam—Verbal	1.43	0.81
Graduate Record Exam—Quantitative	1.47	0.79
Graduate Record Exam—Analytical	1.61	0.96
Law School Admission Test	1.55	1.62
	Minorities[b]	
Medical College Admission Test		
Verbal	1.01	
Quantitative	1.01	
Information	1.00	
Science	1.27	

[a]From statement submitted by Educational Testing Service (Princeton, N.J.) to the U.S. House of Representatives Subcommittee on Civil Service, in a hearing on May 15, 1979.
[b]Differences here are smaller than those typically found for blacks and larger than those typically found for Chicanos, reflecting the fact that the minority data reported here are based on both blacks (N = 2406) and Chicanos (N = 975).

average of the *SDs* of the two groups being compared). The groups taking these tests are all self-selected persons at advanced levels of their education who have already had considerable experience in taking tests in school and presumably understand their reasons for taking these admissions tests. And they surely appreciate the importance of scoring well on them. Hence, it is hard to put much stock in Bernal's claim that minority persons perform less well on tests because they are less sophisticated about them and that they "are being 'put on the spot' to perform like whites on tasks that are of no relevance to them." Is the bar exam of no relevance to a person who has completed 12 years of public school, 4 years of college, and 3 years of law school, and who wants to practice law?

Bernal's "test ambience" theory also seems an inadequate explanation of why some tests show larger white–black differences than others—even tests as similar as the forward and backward digit-span test of the Wechsler Intelligence Scales. The white–black difference is about twice as great (in *SD* units) for backward as for forward digit span, even though both tests are given in close succession in the same "ambience." But backward digit span is more highly correlated with the IQ and the *g* factor than is forward digit span, and this is true within each racial group (Jensen & Figueroa, 1975).

A difference in motivation remains a highly dubious explanation of majority–minority differences. For one thing, there is simply no good evidence for it. In general, motivation, in the sense of making a conscious, voluntary effort to perform well, does not seem to be an important source of variance in IQ. There are paper-and-pencil tests and other performance tasks that do not superficially look very different from some IQ tests and that can be shown to be sensitive to motivational factors, by experimentally varying motivational instructions and incentives, and that show highly reliable individual differences in performance but show no correlation with IQ. And minority groups do not perform differently from whites on these tests. Differences in IQ are not the result of some persons' simply trying harder than others. In fact, there is some indication that, at least under certain conditions, low scorers try harder than high scorers. Ahern and Beatty (1979), measuring the degree of pupillary dilation as an indicator of effort and autonomic arousal when subjects are presented with test problems, found that (a) pupillary dilation was directly related to the level of problem difficulty (as indexed both by the objective complexity of the problem and the percentage of subjects giving the correct answer), and (b) subjects with higher psychometrically measured intelligence showed less pupillary dilation

to problems at any given level of difficulty. (All the subjects were university students.) Ahern and Beatty concluded,

> These results help to clarify the biological basis of psychometrically-defined intelligence. They suggest that more intelligent individuals do not solve a tractable cognitive problem by bringing increased activation, "mental energy" or "mental effort" to bear. On the contrary, these individuals show less task-induced activation in solving a problem of a given level of difficulty. This suggests that individuals differing in intelligence must also differ in the efficiency of those brain processes which mediate the particular cognitive task. (p. 1292)

Bernal's experiment was intended to test his ambience theory. Essentially, four groups of eighth-graders were given two brief cognitive tests (number series and letter series). The groups were white (W), black (B), monolingual English-speaking Mexican-Americans (M1) and bilingual Mexican-Americans (M2). A random half of each group was tested under standard conditions (control), and the other half (experimental) of each group was tested under special conditions of instruction, prior practice on similar test items, and so on, intended to improve test performance. The control groups were tested by a white examiner, the experimental groups by examiners of the same minority ethnic background as the subjects. In addition, Bernal states that the "facilitation condition combined several facilitation strategies designed to educe task-related, problem-solving mental sets that cannot be assumed to occur spontaneously in all subjects . . . and that seem to assist in concept attainment." The exact nature of these "facilitation conditions" is not described. Hence, if they produced significant results, other investigators would be at a loss in their attempts to replicate the study. Whether the experimental treatment was in any way importantly different from those in other studies that have manipulated instructions, coaching, practice, examiner's demeanor, and so on, prior to the actual test, cannot be determined from Bernal's account. But a plethora of other studies in this vein have yielded preponderantly negative results with respect to Bernal's hypothesis, that such facilitating treatment should have a greater advantageous effect on blacks and Mexican-Americans' test performance than on whites' performance.

The results of Bernal's experiment can be seen most easily when presented graphically. Figures 5 and 6 show the mean scores of the four ethnic groups under the experimental and control conditions for the letter series and the number series tests. Figure 7 shows the mean difference (on each test) between the experimental and control conditions for each ethnic group.

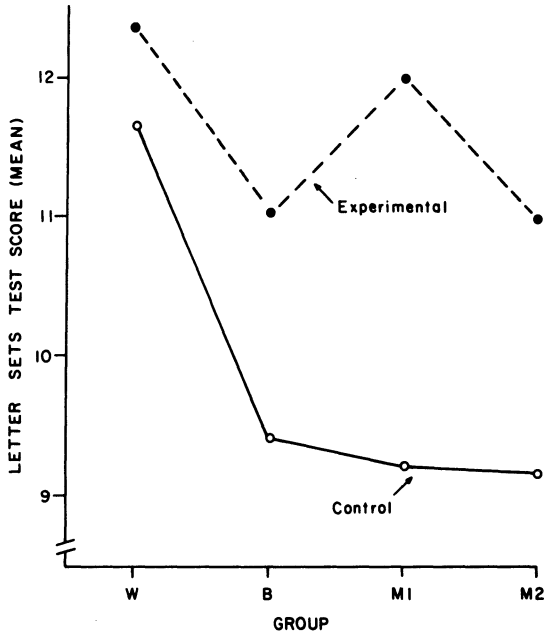

FIGURE 5. Mean scores on letter series test of white (W) and black (B), English-speaking (M_1) and bilingual (M_2) Mexican-Americans, under the control (standard test instructions) and experimental (facilitating pretest experience) conditions.

Bernal commits an unfortunately rather common error in statistical logic in interpreting his results.[1] It has been termed a *Type III error:* testing an inappropriate hypothesis that is mistaken for the intended one. Bernal performed ANOVA separately on the control condition and found that the ethnic groups differed significantly ($p = .04$ for the letter series and $p = .006$ for the number series). Then, he did an ANOVA separately on the experimental condition and found that the ethnic groups did not differ significantly ($p = .483$ for the letter series and $p = .24$ for the number series). He then concluded that his "ambience" hypothesis is substantiated because the four ethnic groups differed significantly under the standard test administration condition and differed nonsignificantly under the

[1]Three months or so before writing this commentary, I personally spoke to Dr. Bernal about this statistical faux pas and suggested that he might wish to emend his paper accordingly, so that I wouldn't have to devote any of my commentary to criticizing his analysis on this point. I have since received no communication about this matter from Dr. Bernal. I have asked the editors to solicit a reply to my comments from Dr. Bernal, to be appended to this chapter.

test-facilitating condition. But this reasoning is a Type III error—an error in statistical logic—because it does not provide a test of the essential question: Did the ethnic groups differ significantly in the difference between the experimental and the control conditions? That is, do the data points in Figure 7 differ significantly among the ethnic groups?

Of what interest is the hypothesis that the significance level of the difference between ethnic groups under the control condition is different from that under the experimental condition? If that were really the hypothesis of interest, then we should be presented with a significance test of the difference between the *p* values for the ethnic groups' main effect under the experimental (E) and control (C) conditions. But that is not the question we want to have answered. What we really want to know is whether the experimental treatment had significantly different effects (i.e., Experiment−Control; E-C) on the various ethnic groups.

Fortunately, Bernal, apparently unknowingly, provides the proper test of this hypothesis in the ANOVAs of his Tables 5 and 6, in which the interaction of treatment × race (A × B in Bernal's tables) is the proper test of the hypothesis. He notes, correctly, that

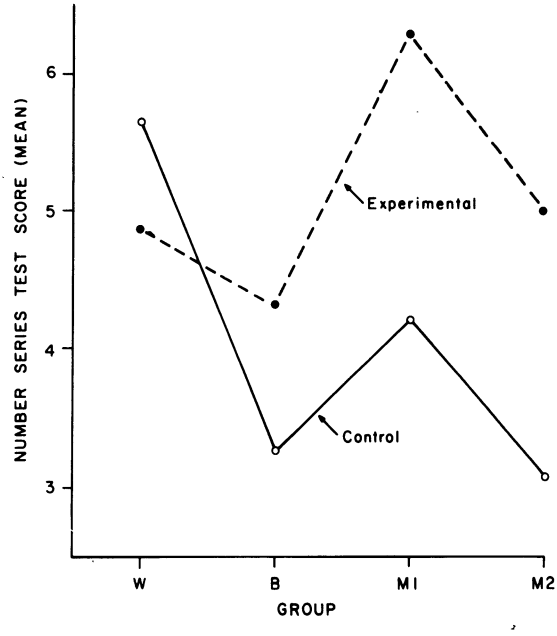

FIGURE 6. Mean scores on number series test.

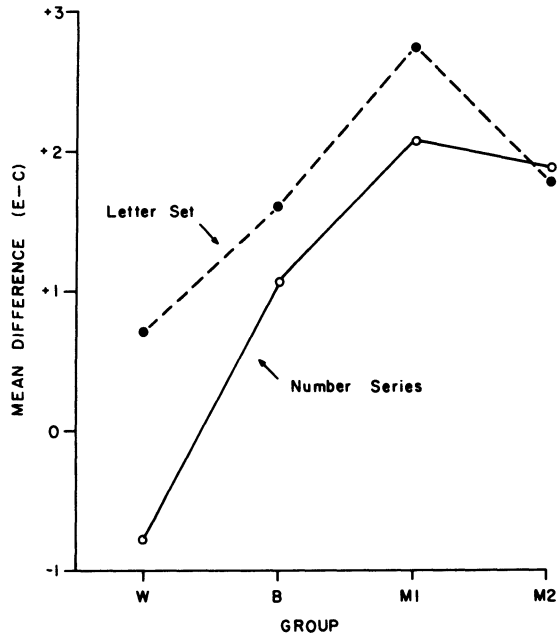

FIGURE 7. The mean difference between the experimental and the control conditions for each ethnic group on the letter series and number series tests. The differences between ethnic groups on the E-C differences (for each test) are what constitute the group × treatment interaction. It is nonsignificant for both the letter series ($F = .67$, $df = 3,168$, $p = .575$) and the number series ($F = 2.17$, $df = 3,168$, $p = .092$).

these interactions are nonsignificant by the usual criteria (for the number series, $p = .092$; for the letter series, $p = .575$). (*Post hoc* Scheffé tests of the contrasts between the E-C difference for the white group and the overall mean E-C difference of the three minority groups are, of course, also nonsignificant for both the letter series and the number series. In other words, the effect of the treatment was not significantly greater for the minorities than for the whites.) The smaller p of .092 for the number series, as we can see in Figure 7, depends mainly on the anomalous condition that the treatment effect resulted in *lower* test scores in the white group. It also seems unexpected that the M_1 group (monolingual English-speaking Mexican-Americans) showed a greater enhancement of test scores by the treatment than did the bilingual Mexican-Americans (Group M_2). Of course, the similarity in configuration of the group mean E-C differences shown in Figure 7 for the two tests does not carry the same significance as if the two tests were based on independent groups.

Because the same subjects took both tests, which were undoubtedly correlated, sampling errors would produce similar profiles of group means for both tests. Because both tests were intended to measure the common factor found in intelligence tests, it would have been a stronger design to have combined the two test scores (after conversion to standard scores) for each subject. This procedure would have minimized test-specific variance and maximized (for the given data) the common-factor variance, making the results potentially of more general interest. Considering the unimpressive significance levels of the group × treatment interactions for the separate tests, however, it is unlikely that the combined scores would appreciably enhance the significance of the interaction.

In summary, the result of Bernal's experiment, when correctly interpreted, does not statistically substantiate his "test ambience" hypothesis; instead, it is quite in line with the preponderance of many other experiments in the same vein, which have similarly yielded nonsignificant treatment × race (and treatment × social class) interactions (see Jensen, 1980a, Chap. 12).

Eysenck

Eysenck has presented a comprehensive and well-balanced review of the main lines of contemporary thinking and empirical evidence bearing on the causes of the observed differences in mental test scores (and all their socially important correlates) among various populations. As I find practically nothing in Eysenck's presentation to which I would take exception, and as I have fully spelled out my own views in this area in my latest book (Jensen, 1981), I will here comment only on a point that seems perpetually to confuse many readers of this literature and to which Eysenck does not point with sufficient warning. It all falls under the heading of what I once labeled the *sociologist's fallacy* (Jensen, 1973, p. 235) because I came across it most often in the writings of sociologists. As I point out in my comments on Mercer's work, the sociologist's fallacy seems to be one of the main pillars of her advocacy of tests with pluralistic norms.

In its simplest form, the sociologist's fallacy consists of attributing an exclusively causal role to socioeconomic status (SES). SES is usually indexed by a host of variables, in some weighted combination, such as occupational prestige; amount and sources of income; amount of formal education; size, condition, and neighborhood of the home; reading material and other amenities in the home; and

membership in civic, cultural, and social organizations. These indices are all highly intercorrelated, so the measurement of any one of them pulls along with it all the others to a large extent.

Eysenck points out that many studies show that when blacks and whites are equated on one of the standard composite indices of SES, the mean black–white IQ difference is generally reduced by something like one-third of a standard deviation. This assertion is factually true. But then readers should immediately beware of making any causal inference, lest they fall into the sociologist's fallacy. For unless we already know that SES is one of the causal factors in IQ variance and that IQ is *not* a causal factor in SES variance, then the interpretation of the reduction in the black–white IQ difference when the groups are equated (either by direct matching or by statistical regression) on SES is at risk for the sociologist's fallacy. Without the prior knowledge mentioned above, the reduction in IQ difference must be interpreted as the maximum IQ difference between the races that could be attributed to the causal effect of the fact that the races differ, on average, in SES. It is logically possible that equating the racial groups on SES could reduce the IQ difference to zero, and yet, not one bit of the IQ difference would be causally attributed to SES. As Eysenck points out, the one inference that we are logically justified in drawing from IQ studies that equate blacks and whites (or any other groups) on SES is that the reduction in the mean IQ difference (generally 3–5 IQ points of the 15-point overall mean black–white difference) is the largest part of the race difference that could be causally attributed to all the variables subsumed in the SES index. The evidence thus clearly shows that the race difference in IQ cannot be explained entirely in terms of the SES difference. And because we know from other evidence[2] that, within each race, IQ has a stronger causal relationship to SES than SES has to IQ, whatever reduction in the black–white IQ difference results from equating the two groups on SES is a considerable overestimate of the effect of SES on the IQ difference.

The few simple path diagrams in Figure 8 shows the most obvious possibilites for the causal connections among race, SES, and

[2]For example, the correlation between individuals' IQs and the SES of the parental homes in which the individuals are reared is much lower than the correlation between individuals' IQs and the SES that they themselves attain as adults. Also, on average, persons who are brighter than their parents (or siblings) attain higher SES as adults than that of their parents (or siblings). Thus, high IQ is causally related to upward SES mobility, and low IQ is causally related to downward SES mobility.

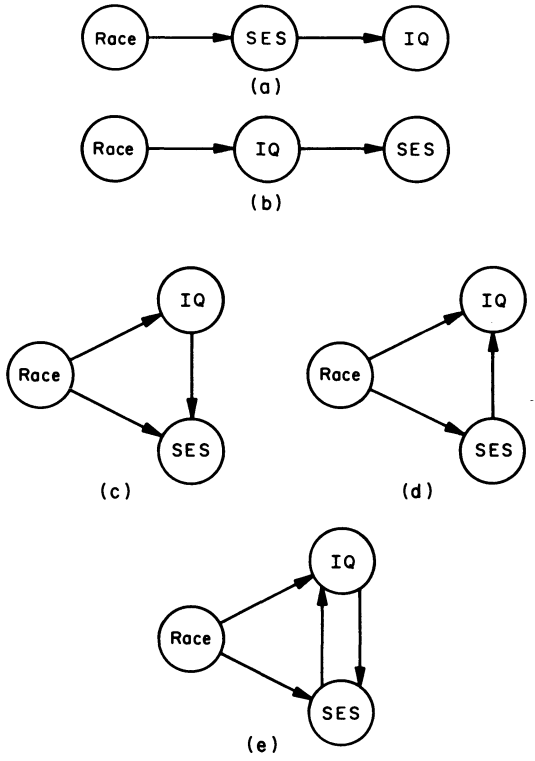

FIGURE 8. Path models illustrating possible forms of causal connections (arrows) among race, social class (SES), and IQ.

IQ. There is no implication in any of these path models that any of these three variables is exclusively the cause of any of the others. IQ and SES are multiply determined by many factors, both environmental and genetic. In each model, the arrows represent the direction of the causal connections between variables.

Model (a) is the implicit assumption underlying all the varied manifestations of the sociologist's fallacy. If this model were indeed correct, we would be justified in matching races on SES when comparing their IQs, or in partialing SES out of any correlations between race and IQ. But this model is clearly contradicted by evidence that shows that there is not just a one-way causal connection going from SES to IQ. The arrow from race to SES in this model would be due to racial discrimination, unequal educational and employment opportunities, and any other racial-cultural factors (other than IQ) that might affect social mobility. Readers should especially note that the

genetic question *per se* is not addressed by any of these models. *Race* in these models denotes all characteristics associated with race, except the variables subsumed under the SES index, regardless of whether they are genetic or cultural in origin.

Model (b) has been much less frequently considered than Model (a), and there is much less good evidence pertaining to it, other than correlational data.

Model (c) seems more realistic than Model (a) or (b), but it is probably too simple in omitting any causality from SES to IQ, although the extent of that causality is not at all well established by empirical evidence. It was on this very point that the necessity of discarding Burt's (1921) questionable (probably fraudulent) "data" on monozygotic twins reared apart constituted the greatest loss to our knowledge on this matter.

Model (d) is inadequate because there is good evidence that, in adults, SES attainment is caused to some extent by IQ.

Model (e) is probably the most appropriate, as it expresses all of the empirically known and plausible relationships, particularly the two-way interaction between IQ and SES. Perhaps, someone will collect all the relevant empirical data and, using the method of path analysis, determine which of these several models (or possibly others) shows the best fit to all the data.

Eysenck mentions still another approach to the study of the connection between race and SES with respect to mental ability differences. This involves an examination of the *profile* of race differences and of SES differences across a number of tests of various abilities. If the profiles of race and SES differences are dissimilar, this dissimilarity is strong evidence that the causal factors in the race difference are not the same as the causes of the SES differences observed within each race.

In this connection, I refer to the study by Reynolds and Jensen (1980; 1983) discussed by Eysenck. Blacks and whites (270 of each) from the national standardization sample of the WISC-R were perfectly matched on full-scale IQ, and a comparison was made from the profiles of the two groups on the 12 subscales of the WISC-R. It was found that whites and blacks differed significantly on certain subscales, even when they were perfectly equated on overall IQ. As the white subjects were specially selected to match the black subjects in IQ, and the mean of the white IQ distribution is about one standard deviation higher than the black mean, we are faced with possible regression artifacts in the profile of subtest scores of the selected white group. As these subjects were selected largely from the lower

half of the white IQ distribution, their scores on the 12 subscales would be expected to regress upward (toward the white mean) by varying amounts, depending on the reliability of the subscales and on their degree of correlation with the full-scale IQ. Therefore, the method used in this study, based on matched groups, could produce results that are simply artifactual if the regression effects are large enough.

Fortunately, there is a better method for comparing the black and white subtest profiles when the groups are, in effect, equated on full-scale IQ. Reynolds and I have now applied this method, using the *entire* national standardization sample of 1,868 whites and 305 blacks (Jensen & Reynolds, 1982). The sampling method for obtaining these groups ensures that they are highly representative of the white and black populations in the United States.

We used the point-biserial correlation (r_{pb}) as the measure of the average white–black difference. (Whites are coded 1, blacks are coded 0, in computing the point-biserial r, so that a *positive* r_{pb} indicates that whites score *higher* than blacks, and a *negative* r_{pb} indicates that whites score *lower* than blacks). The r_{pb} has a perfect monotonic relationship to the mean group difference expressed in standard score units, and within the range of mean differences found in this study, the relationship between r_{pb} and the mean difference is almost perfectly linear, so the relative differences among the various subtests are not distorted by the r_{pb} scale as an index of the racial difference. To show the profile of racial differences when the groups are equated on full-scale IQ (FSIQ), one simply partials out the FSIQ from the race $\times$ subscale r_{ps}. Figure 9 shows the results of this analysis. We see that partialing out full-scale IQ reduced most of the point-biserial correlations between race and subtests to near zero; but with such a large number of subjects, five of the partial correlations were significant at the .05 level (indicated by asterisks). When whites and blacks were statistically equated for FSIQ, the whites significantly exceeded blacks on Comprehension, Block Designs, Object Assembly, and Mazes. The latter three subtests (BD, OA, and M) appear to represent a spatial visualization factor. (Other studies, too, have shown that blacks perform relatively poorly on spatial ability tests, which are invariably the low points in the average ability profiles of blacks.) The difference on the Comprehension test cannot be attributed to the *g* factor (which was partialed out via the FSIQ) or to a verbal factor *per se*, as three other tests that are highly loaded on the verbal factor showed negligible differences. In fact, the best measure of the verbal factor, Vocabulary, showed zero difference between IQ-equated whites and blacks. When equated with the

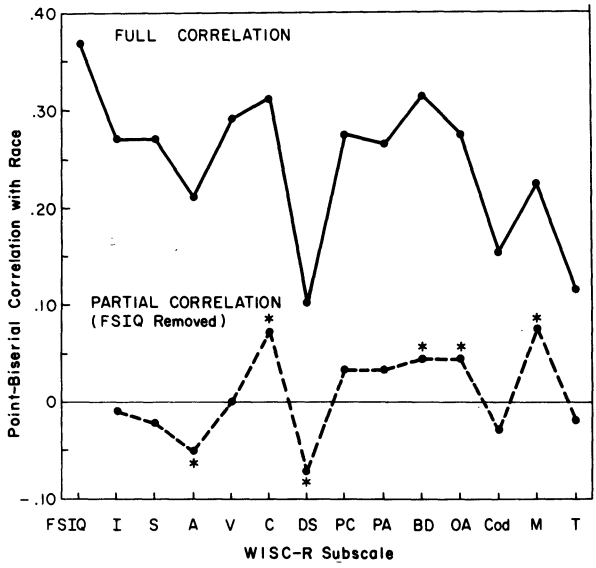

FIGURE 9. Point-biserial correlation as an index of white-black mean difference on full-scale IQ and on each of 13 subtests of the WISC-R (Wechsler Intelligence Scale for Children-Revised). The upper profile shows the actual group differences. (All are statistically significant.) The lower profile shows the white-black differences on the 13 subtests after full-scale IQ has been partialed out, in effect equating the racial groups on general intelligence. Those differences that are significant beyond the .05 level are indicated by asterisks. (I—Information, S—Similarities, A—Arithmetic, V—Vocabulary, C—Comprehension, DS—Digit Span, PC—Picture Completion, PA—Picture Arrangement, BD—Block Designs, OA—Object Assembly, Cod—Coding [Digit Symbol], M—Mazes, T—Tapping [Knox Cubes].)

whites for IQ, the blacks performed significantly better than the whites on Arithmetic and Digit Span. These subtests, along with Coding and Tapping (on which blacks also excelled) are the only WISC-R tests that are loaded on a short-term memory factor, which can be classed as a Level I ability (Jensen, 1974; Jensen & Figueroa, 1975; Vernon, 1981).

The profile of the partial correlations is correlated .96 with the profile of mean differences (in standard score units) obtained by direct matching of whites and blacks in the analysis by Reynolds and Jensen (1980). Apparently, the possible regression effects due to matching subjects from populations with different means did not result in much, if any, distortion of the white-black differences in subtest profiles.

The same correlation analysis (using Pearson *r*) was performed with respect to SES (rated on a 5-point scale) in the white sample (N = 1,895). That is, full-scale IQ was partialed out of the correlations between SES and each of the subtests. The profile of partial correlations (indicating the size of the SES difference on each of the subtests) looks quite different from the corresponding profile for white–black differences. The correlation between the SES and race profiles is *negative:* −.45. In other words, the pattern of ability differences between whites and blacks was quite different—almost the opposite—from the pattern of differences associated with SES. The black–white differences, therefore, cannot be interpreted as an SES difference.

The pattern of nonpartialed correlations (representing white–black differences) in Figure 9 is also relevant to the Spearman hypothesis, which states that the magnitudes of white–black differences on various tests are directly related to the *g* loadings on the tests (Spearman, 1927, p. 379). The WISC-R for the total sample (N = 2,173) was subjected to a hierarchical factor analysis, using the Schmid–Leiman (1957) procedure, to yield a second-order *g* factor. As a test of the Spearman hypothesis, the *g* loadings of the 13 subtests were correlated with the profile of r_{pb}'s (the upper profile in Figure 9), giving a Pearson *r* of + .76, df = 12, $p < .01$, which more or less bears out Spearman's hypothesis. But in evaluating this correlation of +.76, one must take into account the *profile reliabilities* of the *g* loadings and of the white–black differences (r_{pb}). This was done by splitting the sample randomly in half and performing the same analysis separately in both halves. When the profile reliabilities are taken into account, so as to correct the correlation between *g* loadings and white–black differences for attenuation, the corrected correlation is +.84. Thus, the Spearman hypothesis is clearly supported, at least in the sense that *g* is the most important factor in the race difference. But it is not the only factor contributing to the difference. As indicated by the analysis in Figure 9, the groups also differed, albeit slightly, on other ability factors independent of *g*, particularly spatial ability (in favor of whites) and short-term memory (in favor of blacks).

HARRINGTON

Before looking at Harrington's major thesis, a number of side issues raised in his chapter call for comment. Some are misleading.

Harrington states, "A test is considered biased if and only if the effects of such interaction [i.e., group × item interaction] lead to

group differences in means, in predictive validities, or in standard errors of estimate of the test (Jensen, 1980)" (Chap. 3, p. 110). The reference to *Bias* would make it appear that this is a paraphrase of something I have said, but the fact is, I have never said anything of the kind and, in fact, have said exactly the opposite. What I have said in *Bias* is as follows:

> It should be kept in mind that a significant and large group × items interaction can exist even though the groups do not differ at all in their overall mean test score. This means that, according to the criterion of a group × items interaction, a test may be markedly biased without there being an iota of difference between the group means. (p. 435) A test in which item biases with respect to different subpopulations are "balanced out" is still regarded as a biased test from a psychometric standpoint. (p. 455)

The determination of test bias does not hinge on whether two (or more) subpopulations do or do not show significant differences in mean test scores. If a test does not behave the same *internally* in the various groups—that is, if there are differences in the reliability, factor structure, and rank order of item difficulties (i.e., group × item interaction)—then the test's construct validity is suspect as an unbiased measure of ability across the various groups.

Harrington describes hypothetical methods by which test items *could* be selected that would ensure creation of a biased test, in terms of the several standard criteria of bias explicated in *Bias in Mental Testing* (Chap. 9). No one questions that biased tests *can* be constructed, if one wishes to work at it. But that is not how real tests are actually constructed. If one believes that the mean white–black difference on virtually all cognitive tests is the result of bias in item selection, it should be possible to demonstrate that it is possible to *reverse* the direction of the white–black difference in mean test scores by making some other biased selection from a pool of items that measure *g*, that is, that involve some form of relation eduction, for that is the essence of the *g* factor common to all cognitive tests, and it is the largest factor in the white–black difference. No one has claimed that whites and blacks differ, or differ in one direction only, on all conceivable behavioral measurements. The claim is made only about intelligence or cognitive ability tests insofar as these are loaded on the *g* factor common to the vast majority of such tests. So far, no one has been able to devise a cognitive test that reverses the white–black difference, despite numerous intensive efforts to do so (reviewed in *Bias*). The so-called "chitlin" test of knowledge of black ghetto slang, or the very similar Black Intelligence Test, or the BITCH

test, does not qualify as a successful attempt in this vein, as none of them has demonstrated construct validity of any kind, or factorial validity for any mental abilities, or predictive validity for any practical criteria. At most, these "tests" can claim only face validity, as measures of knowledge of black ghetto argot, and there is not even evidence that they are psychometrically adequate measures of that. Scores on these tests are most probably *negatively* correlated with the kind of upward socioeconomic mobility that is the proclaimed goal of the black leadership in the United States.

It would be surprising, of course, if these tests of black argot, like all vocabulary tests, did not also have some *g* loading within the group that uses this particular argot. But the extreme subcultural specificity of such tests makes them unsuitable as measures of any broad cognitive abilities, such as *g* and verbal ability, even among black Americans. The BITCH test, for example, has shown correlations of −.04 with the Wechsler Adult Intelligence Scale (WAIS) Verbal IQ, +.13 with Performance IQ, +.04 with full-scale IQ, and a correlation of −.33 with level of education in a black sample averaging two and one-half years of college (Matarazzo & Wiens, 1977).

Harrington states: "It is possible that whatever intelligence tests measure is devoid of evolutionary significance or survival value" (Chap. 3, p. 130). Yes, "possible," but most *improbable*. The one feature that most distinguishes the human species from the rest of the animal kingdom is humans' superior intelligence, made possible by the biological evolution of a large and complex cerebrum. The cerebrum has tripled in size in the course of human evolution, despite the fact that there are many anatomical and perinatal disadvantages to increased brain and cranial size. The only selective advantage in the evolution of greater brain size is the more complex behavioral capacities that it confers. The greatest development of the brain has been of the neocortex, especially those areas serving speech and manipulation. Tools found with fossil remains indicate that increasing brain size was accompanied by the increasing complexity of tools, and along with the development of complex tools are also found artistic drawings on the walls of caves. In the last 1 or 2 million years, the strongest selection pressure in humans has been for behavioral traits of increasing complexity, accompanied by the increasing size and complexity of the cerebrum. Konrad Lorenz (1973), the first behavioral scientist to win a Nobel prize, has expounded the thesis that the evolution of the complex functions of the human brain that make possible such intelligent operations as comparing, analyzing, separating, seeing relationships, classifying,

counting, abstracting, conceptualizing, recalling, imagining, planning, and the like came about from selection by environmental demands acting directly on the behaviors made possible by increasingly complex nervous functions. These are the behavioral capacities that, in large part, are measured by our present tests of mental ability, in which the largest unidimensional part of the individual differences variance is termed the g factor. Independent evidence that our present most g-loaded tests tap an ability that has undergone directional selection in the course of human evolution is the presence of genetic dominance deviation, revealed in quantitative-genetic analyses of test data. Dominance effects are revealed especially in studies of inbreeding depression, which is found for IQ as well as for certain physical traits. Stature, for example, has also increased in the course of evolution and also shows genetic dominance (Jensen, 1978). Genetic dominance (and other nonadditive effects of genes) increases, as a result of selection, for those traits that are advantageous in the species' struggle for survival.

Harrington suggests that the reason that the observed racial group differences on tests remain even when a test shows identical predictive validity for both racial groups is that the measurement of the criterion is itself as biased as the predictor test. One explanation for bias in the criterion measurement, he hypothesizes, is that the same biased test-item selection procedures that are used in the construction of the predictor test are also used in the construction of the criterion test, making it equally biased.

There is nothing we know that would *a priori* rule out this possibility for some particular rare tests and the criteria on which their validity is based. But I believe that the hypothesis is of very limited generality and cannot be accepted as an explanation of the typical majority–minority differences in test scores or the typical finding that the test scores are unbiased predictors of educational and occupational criteria. First of all, there is no general evidence (aside from Harrington's experiment) that the usual item-selection procedures used in test construction will automatically bias a test against a minority group. Second, not all validity studies are based on correlating a predictor test with a similarly constructed criterion test (e.g., a standardized scholastic achievement test). The criterion measure is often an actual work sample, an objective productivity measure, performance ratings, or course grades. Often, when the criterion measure is a test score, it is from an informally constructed test, such as most teacher-made tests, examinations in college courses, and specific job-knowledge tests—none of them based on the classical tech-

niques of item selection. Yet, scores on these tests and all the other types of criterion measures also show majority–minority differences and are predicted with equal validity in majority and minority groups by standard aptitude tests. The tremendous variety of criteria that are predicted by aptitude tests and the many hundreds of studies that have failed to show significant differential validity for whites, blacks, and Hispanics render highly improbable the hypothesis that congruent biases in predictor tests and criteria account for the group differences and the absence of differential validity.

Harrington's main thesis is based on an interesting experiment with six genetically different inbred strains of laboratory rats. They were given multiple learning trials in a set of mazes with varied stimulus attributes and problem configurations in which the rats' performance was scoreable, analogous to the items of a psychometric test. From the total set of all possible scoreable units ("items") of learning performance in these mazes, tests were made up by selecting a subset of "items," by the use of one of the classical selection criteria in psychometric practice: the item–test correlation. Different proportions of the various genetic strains of rats were included in the "standardization sample" on which the subset of "items" was selected. It was then found that the mean test scores (based on the selected subset of "items") differed across the various strains. The mean scores were directly related to the proportional representation of each strain in the "standardization sample." From these results of the experiment, Harrington has drawn the following generalizations:

> First, the mean performance of homogeneous groups on tests tends to vary directly with the extent of representation of the groups in the population used for psychometric construction of the tests.
>
> Second, the predictive validity of tests for members of homogeneous groups tends to vary directly with representation of the groups in the population used for psychometric construction of the tests. . . .
>
> The data of this research program provide one explanation of minority group differences in test performance. The results are applicable to all forms of tests. They imply a general tendency for tests to be biased against minorities and to have less validity when used with minorities. (Chap. 3, p. 134)

I have no complaint with Harrington's effort as a contribution to experimental behavioral genetics. The result is indeed interesting. But it seems to me that it merely poses a problem. It does not answer any question about human minority-group test-performance. What needs to be explained is why, with these particular inbred strains of rats, one finds the interesting phenomenon described above, which,

for brevity, I shall dub the *Harrington effect*. And why are data on humans so lacking in evidence of this phenomenon?

The Harrington effect is interesting in its own right, as a demonstration of genetic differences in the factors involved in maze learning in strains of rats. But I see no justification or need to generalize the conclusions from strains of rats to races of humans—the species of practical concern with regard to test bias. We already have much direct evidence, based on humans, that the Harrington effect cannot be generalized to human racial differences in test performance (e.g., see Reynolds, 1982). No amount of experimentation with *rats* can possibly nullify all the evidence based on *human* test results that goes directly contrary to the generalizations from Harrington's rat experiment.

For example, Asians and Jews are minorities in America that score as high as or higher than the majority on majority-standardized IQ tests and college entrance exams. Japanese in Japan, on the average, outperform American whites on the U.S. norms of the Performance Scale of the Wechsler IQ test. Arctic Eskimos perform on a par with British and American norms on the British-standardized Raven's Matrices Test. We have recently discovered that Chinese children in Hong Kong outperform white children of the same age in California on the Raven. African infants score higher on the American-standardized Bayley Infant Scales of Development than do the middle-class white American infants on whom the test was originally developed.

Still more direct counterevidence to the Harrington effect in human populations is found in the application of one of the methods described in *Bias* (pp. 580–583) for detecting biased items: the item selection method. Subtests are made up by selecting items from a large pool of items according to the usual psychometric criteria for item selection, on samples from the two (or more) subpopulations in question. The method was applied separately to large samples of blacks and whites by Green and Draper (see Jensen, 1980a, pp. 581–583), creating somewhat different subtests, each derived by the same item selection procedures in the different racial groups. The items thus selected in each subtest are the "best" selection of items for each group, according to common psychometric criteria. This procedure would seem to provide a direct test of Harrington's hypothesis on human subjects. When the two subtests were given to blacks and whites, the average white–black difference (in standard deviation units) on the white-derived subtest was 0.78σ; it was 0.85σ on the

black-derived subtest. (Both differences are in favor of whites.) The authors of this study concluded:

> The amount of relative improvement in score that a minority group could expect to gain by using tests built with tryout groups like itself does not appear to be very large. The relative improvement is most unlikely to overcome any large discrepancy between typical test scores in that group and those in more favored groups. (Green & Draper, 1972, p. 13)

Another direct test of Harrington's hypothesis was the construction of the Listening Comprehension Test (LCT) by the American Institutes of Research (which is fully described in Jensen, 1980a, pp. 678–679). The LCT was devised completely within a low-SES black population, following all of the usual psychometric procedures of test construction. After the test was developed entirely on blacks, it was tried on other samples of blacks and whites of middle and low SES levels. In every comparison, whites scored higher than blacks. Although the test was devised on low-SES blacks, that group scored 1.32σ lower than middle-SES whites. Moreover, the LCT had equally good validity for blacks and whites as a predictor of scores on a standard test of verbal ability. Thus, a number of studies contradict the Harrington hypothesis with human samples. On the other hand, I can find no study in the psychometric literature that affords any support to Harrington's hypothesis. It seems strange that when biological psychologists like Harrington urge such extreme caution about generalizing, say, the results of heritability studies from one human racial group to another, they nevertheless show no hesitation in generalizing experimental results directly from rats to humans!

Finally, Harrington argues that, because he finds little or no evidence of a general ability factor in the maze performance of rats, this lack of evidence somehow brings into question the *g* factor in the test performance of humans. Because the rat behavior appears to be highly "multifactorial," Harrington concludes:

> To suggest that such results are true only for animals and not for humans is to argue that the rat is intellectually a much more complicated creature than is the human being. Yet this, it seems to me, is the implication of the *g* hypothesis. (Chap. 3, p. 130)

Harrington's conclusion, however, is a sheer *non sequitur*. It rests on a confusion of complexity of mental processes with factorial complexity. The factors of factor analysis depend on covariation among various tests of abilities. In principle, there is no necessary connection between the complexity of the cognitive processes involved in any of the tests and the degree of covariance among the tests—the

covariance that gives rise to factors. Whether there is any connection between the cognitive complexity of the processes required by the various tests entering into a factor analysis and the degree of simplicity or complexity (i.e., the number of factors) of the emergent factor structure is a strictly empirical question. The number of factors emerging from *n* tests of abilities carries no necessary or logical implication about the complexity of behaviors or inferred cognitive functions involved in the test performance. However, as pointed out in Chapter 6 of *Bias,* there is now considerable evidence, from human test data, that the factor analysis of tests involving relatively more cognitive complexity yields a smaller ratio of factors to tests than the factor analysis of relatively simple tests. Clark L. Hull (1928) made this important observation more than half a century ago, in factor-analyzing a large collection of tests including tests of sensorimotor skills, coordination, reaction time, rhythm, balancing, memory, tapping, card sorting, and verbal and nonverbal intelligence tests of various types. He concluded that

> The highly complex intellectual activities correlate highly with each other, the less complex correlate with each other to an intermediate degree, and the relatively simple motor activities correlate with each other only slightly. (p. 215)

Many more recent studies fully substantiate Hull's observation (see *Bias,* pp. 213–222; 229–233), and it is the *g* factor that is loaded most heavily in the relatively more complex tests, especially those calling for some form of relation eduction, as Spearman (1927) noted. Thus, a very large *g* factor and relatively large, but few, group factors, in addition to *g*, are empirically associated with the greater cognitive complexity involved in the tests subjected to factor analysis. The highly multifactorial nature of the behaviors that Harrington noted in his rats' maze-learning activity is much what one would expect from the factor analysis of relatively simple sensorimotor abilities, even in humans. Thus, Harrington's interesting finding is not at all in conflict with the *g* theory of intelligence and even seems to confirm it.

Hilliard

Like all the other critics who have disliked *Bias in Mental Testing,* Hilliard steers clear of the book's main findings and conclusions. Instead of challenging these, he takes up a number of side issues and alludes to supposedly germane research that the book failed to

include. But it must be outstandingly apparent to readers that Hilliard never summons any empirical evidence or closely reasoned arguments based thereon that would support a position on test bias contrary to that expounded in my book. He writes as though there exists some body of evidence that would comfort those who dislike my book's conclusions, but that I have chosen to ignore, and that, if I properly considered it, would overturn the book's main conclusions based on the massive evidence reviewed in the book. But he does not tell us what this contrary evidence is or how it is supposed to contradict all the other evidence that has been brought to bear on the issue of test bias. Hilliard characterizes my book as a review of "highly selected empirical research," whereas, in fact, I tried my best to include everything I could find in the research literature on test bias, and certainly nothing else comes near my book in comprehensiveness on this topic. It was never intended to be, as Hilliard claims, "an exhaustive review of all the relevant literature that pertains to the IQ argument." And certainly, neither I nor anyone else, to my knowledge, has ever had the fatuous thought that it is "the book for the century," as Hilliard suggests. *Bias* was intentionally quite narrowly focused on those aspects of psychometric theory and research most relevant to the problem of test bias. (However, my latest book—Jensen, 1981—gives a much more comprehensive overview of what Hilliard terms the "IQ argument.") Of course, anyone who wishes to argue that my coverage of the research on test bias is itself a biased selection (which I deny) is free to review whatever other evidence bears on the issue and to explain how it would alter the conclusions based on the evidence that I have presented. In view of Hilliard's claim, we might reasonably have expected him to do just that in his present chapter. But neither Hilliard nor any other critic of *Bias,* out of some 100 published critiques, so far, has attempted to do so. I suspect they would if they could. There has not been a scarcity of ideologically or emotionally hostile criticisms, but all of it is substantively innocuous.

Hilliard seems to be arguing that cultural differences between blacks and whites are chiefly responsible for the typical white–black differences in test scores. But as I have pointed out earlier in this chapter, the results of our studies of black and white test performances, at every level of psychometric analysis, from single test items to broad factors, indicate that the cultural differences between whites and blacks in the present-day United States have been grossly exaggerated by those who would insist on a purely cultural explanation of the racial difference in test performance. Analyses of test

data in terms of both internal and external criteria of bias yield results that are quite incompatible with the hypothesis of large cultural or experiental differences between blacks and whites, at least as these affect test performance. I will not belabor this point further. The evidence is there for all to see (*Bias,* especially Chaps. 10 and 11).

Hilliard emphasizes linguistic differences as a chief source of cultural bias and argues that the field of linguistics somehow contains the antidote to what he views as the current unpalatable psychometric conclusions about the absence of cultural bias in widely used standardized tests. But theoretical psychometricians as well as pragmatic users of tests are quite unimpressed by the linguistic arguments, in view of the well-established finding that the black deficit is no greater on tests of verbal abilities than on wholly nonverbal and performance tests. When the general factor is extracted from a large and diverse battery of verbal and nonverbal tests, we find that blacks and whites differ almost entirely on the *g* factor and not at all on the verbal factor, least of all on vocabulary, after *g* is partialed out (see Figure 9). Chapter 9 of *Bias* reviews the many attempts to vary the familiarity of the test contents, with the consistent result that the white–black differences remain fully constant across all these test variations. Only tests of rote memory and motor skills show negligible differences. On all other types of tests, the race differences are substantial. But when *g* is partialled out, hardly any test shows an appreciable difference between blacks and whites. The racial difference is a difference in *g* and not just a linguistic difference or a difference dependent on any special type of item content. Hilliard comes down especially hard on vocabulary tests, as they would appear to be the most quintessentially cultural of all test types. But the lower scores of blacks on the vocabulary subtests of standard scales such as the Stanford–Binet and the Wechsler probably do not underestimate black children's functional vocabulary, whether estimated by their use of standard English or of their own patios—what Hilliard would call the "normal vocabulary" of a particular cultural group. In a study by langauge experts in Detroit, tape recordings were made of black children's speech, and it was discovered that their vocabulary contains only about half as many words as white children's (Silberman, 1964, p. 283). A comprehensive review of research pertaining to the cultural-linguistic hypothesis of the black IQ deficit concluded there was no evidence that supports it and that the explanation of the black IQ deficit must be sought elsewhere (Hall & Turner, 1971).

Hilliard objects to a social definition of race instead of a strictly biological criterion and asks how IQ researchers select "black" or "white" samples for comparison. The answer is, of course, that they do it in the same was as those who assess racial balance in the public schools, or the proportions of different racial groups in special classes, or an institution's conformity to federal guidelines for affirmative action, or for trying court cases of racial discrimination. Although there is a high correlation between the ordinary socially recognized categories of races in the United States and strictly biological criteria of classification, involving a host of visible physical characteristics as well as blood groups[3] and biochemical factors, it is only the social and cultural definition of race that is actually relevant to the study of test bias as it concerns all the practical uses of tests. Moreover, if the observed test-score differences between racial groups are due only to social-cultural factors, as Hilliard claims, then the social definition of race should be quite adequate and, in fact, should be the only appropriate definition. If it is argued that two socially defined racial groups that differ in mean IQ are not racially "pure," by strictly biological criteria, and that one or both groups have some genetic admixture of the other, it can mean only that the biological racial aspect of the IQ difference, if such exists, has been *under*estimated by comparing socially, rather than genetically, defined racial groups.

The chapter by Lloyd Humphreys should provide adequate background for evaluating Hilliard's claim that intelligence "has no common definition among the community of scholars who study it." In fact, there is high agreement among the experts about what they mean by the term *intelligence*. The issue has been the subject of an empirical investigation. Yale psychologist Robert Sternberg devised an elaborate questionnaire intended to assess people's conceptions of the meaning of *intelligence* and the specific types of behavior that they recognize as instances of whatever they mean by *intelligence*. The questionnaire was sent to a representative sample of Ph.D. psychologists who do research and teach courses in the area of human abilities. The questionnnaire was also given to laypeople. Sternberg,

[3]Today, the average percentage of Caucasian genes in persons who are socially identified as black and who so identify themselves, in America, is estimated, on the basis of blood group analysis, at something close to 25%, with a standard deviation of about 14%. The frequency of genes of African origin among persons socially identified as white is estimated at less than 1%. (A detailed discussion of this research, with complete references, is to be found in Jensen, 1973, Chap. 9.)

Conway, Ketron, and Bernstein (1980) reported a very high degree of concordance between psychologists and laypeople about the meaning of *intelligence*. This remarkably high consensus among experts and laypeople as to the subjective meaning of *intelligence* and the recognition of its behavioral manifestations clearly contradicts the notion that *intelligence* is an esoteric technical concept or that there is little agreement among persons concerning its manifest characteristics.

How valid is Hilliard's claim that IQ tests differ widely and bear merely an association with each other? It is granted that the specific item content differs greatly among IQ tests. But the truly remarkable fact is that despite the great variation in types of item content, all IQ tests measure much the same mental factors, especially the *g* factor, which is predominant in all such tests. To get some idea of how widely IQ tests differ, I have determined the average intercorrelation among 30 different published IQ tests, gathered from various studies reported in the literature. I determined the median correlation for each test of all its correlations with other tests. The mean of these median correlations for all 30 tests is .77. But the average reliability of all of the tests is .90, and so we must correct the mean correlation of .77 for attenuation, which brings the true correlation among the tests up to about .86. The median correlation between the Stanford–Binet and the WISC in 47 studies is .80 (or .85 when correlated for attenuation). This is indeed a high degree of agreement among different IQ tests, considering the great variety of samples used in those studies, with widely varying degrees of restrictions of range, and considering the fact that there is some amount of "method variance" among all these tests, which include group paper-and-pencil tests, individual tests, and verbal, nonverbal, and performance tests. The fact that their true-score intercorrelations average about .86 in a wide variety of samples indicates that a large common factor, namely *g*, runs throughout all of these IQ tests. This clearly belies the essence of Hilliard's claim that "IQ tests differ widely."

An important criterion of the absence of test bias, for Hilliard, is evidence that "the same mental process is being measured in two or more cultural groups whose standardized IQ test scores are being compared." One way of examining the test performance of different groups for an answer to this question is by looking at the degree of similarity of the factor structures and factor loadings on the various scorable parts of a test for the two groups in question. C. R. Reynolds and I have recently done this. We subjected the WISC-R national standardization data to a hierarchical factor analysis (Schmid & Leiman,

1957)[4] separately in the white and black samples, with numbers of 1,868 and 305, respectively. (See the chapter by Humphreys for a description of this type of factor analysis, which he deems the most satisfactory method for extracting the general factor, *g*.) We found that both the factor structure and the factor loadings of the 13 subtests of the WISC-R standardization edition (the 12 subtests of the WISC plus a Tapping subtest later deleted from the WISC-R) were virtually identical in the white and black samples, despite the difference of 15.8 points between the group means on the full-scale IQ. The coefficient of congruence (an index of factor similarity, on a sacle from 0 to 1) was computed between blacks and whites for each of the four WISC-R factors: general factor (g) = 1.00, Verbal factor = .99, Performance factor = .98, Memory factor = .98. If Hilliard knows of any bona fide evidence that blacks and whites differ in the types of mental processes that they bring to bear on standard IQ tests, he should bring it to light. We are not aware of any such evidence.

Now, two minor points:

First, the test bias issue does not in the least hinge on settling the question of the true form of the distribution of intelligence in the population. Moreover, I have never claimed that scores on any particular type of test, such as information or vocabulary, should be assumed to have a normal distribution. I have said that many psychologists, for a number of statistical, genetic, biologically analogical, and scientifically heuristic reasons, have explicitly assumed that the latent trait of general intelligence is normally distributed, and that this theoretical assumption is reflected in most IQ scales and derived scores on other cognitive tests standardized on the general population.

Second, Sir Cyril Burt, whom Hilliard refers to as a politician, was never a politician in any sense of the word. In fact, many of his long-time close associates were totally unaware of his very private political views. (His sympathies were with the socialist Labor Party of Britain.) Interestingly, as also noted by Reynolds and Brown in this volume, Burt (1921) was one of the first psychologists to draw attention to the problem of test bias, with respect to social class differences, not long after the publication of the first Binet intelligence scales. (My final conclusions regarding the notorious scandal surrounding Burt's data on identical twins are detailed elsewhere; see Jensen, 1981.)

[4]We are grateful to Professor John Schmid for doing these factor analyses for us.

Finally, the present massive research on our standard mental tests, their associated group differences and all their educationally, occupationally, and socially significant correlates, and the consistent failure to demonstrate by means of any objective evidence that the tests are biased against blacks, constitute an impressive and important body of evidence for psychometric theory and practice. Humphreys (1980b) has summarized the implications very well:

> The measured differences in intelligence are real barriers to equal access by the majority of blacks to higher education, to skilled occupations, and to the professions. The measured differences are causally related to high levels of unemployment and to below average incomes for blacks. The differences and their direct effects are also indirectly related to such social pathologies as higher rates of delinquency and crime in the black population. (p. 55)

To pretend that these conclusions can be likened to the "emperor's new clothes" is, I suspect, only wishful denial—an ineffectual and fatuous response to the reality and the import of the evidence. If there is anything as truly unsubstantial as the "emperor's new clothes" in the IQ cultural bias debate, it is probably the evidence that Hilliard seems to imagine would contradict the main conclusions of *Bias in Mental Testing*.[5] If Hilliard claims to disagree with my definitions of test bias, or with the proposed methods of objectively recognizing bias, or with the empirical evidence on which my conclusions, within this framework, are based, then I think he is obligated to state an alternative definition of bias, to formulate other explicit methods by which one can detect bias, and to cite evidence that specifically contradicts my conclusions. Hilliard has done nothing of the kind. Nor, to my knowledge, has anyone else.

HUMPHREYS

Humphreys's chapter is one of the most lucid and enlightening essays on intelligence that I have come across in my wide reading in this field. It merits thoughtful reading by everyone with an interest in this topic.

[5]Jensen, 1980a, p. 740: "The observed mean differences in test scores between various [racial and social-class] groups are generally not an artifact of the tests themselves, but are attributable to factors that are causally independent of the tests. . . . The present most widely used standardized tests can be used just as effectively for blacks as for whites in all of the usual applications of tests."

The only point on which I have any serious reservations may or may not be a fundamental one—I am not sure. It involves Humphreys's formal definition of intelligence. Not that anyone could sensibly disagree with it, as far as it goes, but it does not go far enough, in my opinion. In one way, it is not sufficiently precise, and in another way, it is not sufficiently open-ended. But before proceeding further, I should restate Humphreys's definition:

> Intelligence is defined as the entire repertoire of acquired skills, knowledge, learning sets, and generalization tendencies considered intellectual in nature that are available at any one period of time. An intelligence test contains items that sample the totality of such acquisitions. . . . The definition of intelligence here proposed would be circular as a function of the use of intellectual if it were not for the fact that there is a consensus among [cognizant] psychologists as to the kinds of behaviors that are labeled intellectual. Thus, the Stanford–Binet and the Wechsler tests can be considered examples of this consensus and define the consensus. (Chap. 7, pp. 243–244)

First of all, there is no hint in this statement that, among all the "repertoire of acquired skills" and so on, some things might be weighted differently from others in the degree to which they represent intelligence. Einstein never knew how to drive an automobile, but he had mastered tensor calculus, an abstruse branch of mathematics. Are we to assign these skills equal (negative and positive) weights in forming a judgment of Einstein's intelligence? I dislike the idea of leaving the relative weights to be assigned to these skills up to the subjective judgment of any one psychologist or any collection of psychologists, cognizant or not. A consensus of expert judgment, it seems to me, is a weak basis for scientific theory. The overthrow of expert consensus is a prominent feature in the history of science. Thanks to Spearman, Burt, Thurstone, and others, we now have a set of tools—namely, factor analysis—for dealing more objectively with the weighting problem. I think that this was the essential contribution of the now-classic paper by Spearman (1904). It seems a reasonable guess that if we included skill in driving an automobile and skill in tensor calculus among the "entire repertoire" referred to by Humphreys, and if we analyzed the whole works, tensor calculus would have a higher *g* loading than automobile driving. Value judgments, academic snobbery, and the like would not enter into this conclusion, as they might do in a mere consensus among Ph.D.'s.

The definition refers to "acquired skills, knowledge," and so on. This definition does not take care of types of behavior that cannot be construed as learned (at least as far as the source of individual differ-

ences is concerned) and that may not be deemed "intellectual" by any expert consensus, but that are nevertheless found to be g-loaded when factor-analyzed along with other types of performance deemed "intellectual." I have in mind, for example, choice reaction time, which is not a learned skill and which under some experimental conditions shows no learning or practice effects whatever; yet, it is correlated with Stanford–Binet and Wechsler IQs (Jensen, 1982b). Clearly, IQ tests measure something *more* than just learned skills and bits of knowledge, although these may serve as adequate vehicles for measuring whatever it is that the test measures, which is something more, and different from, the vehicle itself. If that were not so, why should the verbal subtests of the Wechsler correlate .80 with the performance subtests, with which they have virtually nothing in common at the "phenotypic" level of item contents, knowledge, and skills? No, I think our intelligence-test scores are only the tip of the iceberg.

So, I think we need a deeper conception of intelligence than that suggested by Humphreys's definition; which seems to imply that intelligence consists of no more than what we can readily see with the unaided eye at a given point in time. It also seems to be merely whatever we *say* it is, albeit said by a consensus of the cognoscenti, instead of a wondrous phenomenon, the full nature of which still awaits discovery by scientific means. Everyone knows in general what *universe* means, although astronomers and cosmologists know more about it than laypeople. By rough analogy, Humphreys defines intelligence much as if we said that the universe is simply the myriad specks of light that we all can see when we look up at the sky on any given night. With this approach, what would be the incentive to build a telescope or, if we already had a telescope, to want a more powerful one? When astronomers obtain a more powerful telescope and other instruments for the analysis of starlight, they discover things that were in neither their observations nor their imaginations before. We must presume, if we are not completely solipsistic, that this is so because there is indeed a reality out there, referred to as the *universe,* and that it amounts to a great deal more than our present conceptions of it. It is still being discovered, and there is still much to be understood scientifically about what has already been discovered. I think much the same sort of thing is true of the concept of intelligence. Is there anything in Humphreys's definition, for example, that could have led anyone to expect, much less look for, a correlation between IQ and the frequency and amplitude of evoked electrical potentials in the brain? I believe that there *is* some "nature" under-

lying our observations and test measurements of what we call *intelligence.* To use the word *entity,* as Humphreys does, to describe the "nature" or "reality" underlying our measurements (and factors) is to set up a straw man, if by *entity* he implies some "thing"—a single cause, anatomical structure, physiological mechanism, or biochemical substance.[6] If I really thought that there was nothing more to intelligence than the IQ test scores and more of the same kinds of things we already know about them, I would change my field of research immediately and take up something scientifically more interesting.

Because I do think that there is more to intelligence than merely the behavioral vehicles by which it can be expressed or measured, I think it is theoretically important to retain the ability–achievement distinction, although from that point on, I would agree with everything else that Humphreys says about it.

Humphreys's definition of intelligence seems to be in the tradition of the strictest logical positivism. This philosophy of science, which once had great appeal to me, now seems to me less convincing, and it is my impression that it has already lost favor, generally, in the most advanced physical sciences. Whereas Humphreys insists on an aseptically explicit operational definition of the concept of intelligence, I tend to regard it more as an open-ended theoretical construct still in the process of being explored and more fully understood scientifically. I am reminded of a modern philosopher of science, Yehuda Elkana, who, in the frontispiece of his book on the discovery of the conservation of energy (1974), quoted a statement by H. A. Kramers, which, at several points in his book, he referred to as being of key significance in scientific progress: "In the world of human thought generally and in physical science particularly, the most fruitful concepts are those to which it is impossible to attach a well-defined meaning." I think intelligence is such a concept.

Fortunately, at least from my own standpoint, everything else that Humphreys says in his chapter does not seem to hinge at all on his formal definition of *intelligence.* My own "working definition" of *intelligence* is the general factor of a large and diverse battery of cog-

[6]Elsewhere (Jensen, 1982a), I have written, "It is a mistake to waste time arguing about the definition of *intelligence,* except to make sure everyone understands that the term does not refer to a 'thing,'" and elsewhere, we should heed Miles' (1957) advice: 'The important point is not whether what we measure can appropriately be labelled 'intelligence,' but whether we have discovered something worth measuring. And this is not a matter that can be settled by an appeal to what is or is not the correct use of the word 'intelligent.'"

nitive tests, and this definition does not seem to conflict with anything I find in Humphreys's paper.

Whatever may be the conception of intelligence that actually guides Humphreys's thinking and research in this field, it obviously has not hindered his continuing to make creative and important contributions to our understanding of the nature and the measurement of intelligence, of which we are presented a good sample in his present chapter. We all can learn from it.

Hunter, Schmidt, and Rauschenberger

In their chapter, these investigators present an excellent summary of their many important original contributions to the study of test bias, particularly as it concerns the use of tests in personnel selection. Probably, no one else in the field has done more than these researchers to dispel the twin illusions of situational specificity of test validity and differential validity for majority and minority populations. The type of meta-analysis of the mountains of validity data that led to their conclusions is one of the signal contributions to both the methodology and the substantive knowledge of psychometrics and personnel psychology. In addition, they have highlighted more explicitly and rigorously than has anyone else the practical consequences—in terms of work productivity and dollars and cents—of test validity, various personnel selection models, and minority quotas. These objective cost–benefit analyses may come as a shocking surprise to many of those who would belittle the practical importance of mental tests or who imagine the world would be better off without them.

Manning and Jackson

The chapter by Manning and Jackson is an extremely valuable, empirical, research-based defense of the Scholastic Aptitude Test (SAT) and other advanced educational aptitude tests developed by the Educational Testing Service. Comprehensive and detailed documentation of Manning and Jackson's claim of equal validity of the SAT for white and minority populations in predicting academic performance in college can be found in Breland (1979).

The Educational Testing Service (ETS) is the developer and publisher of the College Entrance Examination Board's SAT, which has practically preempted the field of college aptitude testing. Partly as a result, the ETS in recent years has been assailed by critics from out-

side the psychological and educational testing discipline. The attacks on the ETS have resulted not because anyone has been able to show that their tests are technically substandard or that they are not as valid for racial minorities as for whites, but mainly because the ETS tests are virtually unassailable by these criteria and hence reveal all too clearly unwelcome and disturbing facts about inequalities in the distribution of mental abilities in the nation's population—particularly those developed abilities that are necessary (although not sufficient) for academic attainment beyond high school.

From a strictly psychometric standpoint, the ETS has no real problem countering its critics: High-powered technical expertise and all the statistical evidence are on its side. But I think the ETS repeatedly displays a sorry spectacle in its squirming to offer excuses for the unwelcome social realities that its tests so soundly reveal. We see more examples of it in Manning and Jackson's chapter. But my complaint is not limited to the ETS; other test publishers, in my observation, are in the same straits. Of course, one may easily sympathize with the predicament of a commercial establishment trying to maintain good public relations and avoid controversy. But it seems to me that the ETS and other test publishers have chosen the wrong stance to avoid the heat.

The same good science that is found in the research and test development of the major test publishers should also be evinced in their public statements about socially touchy issues, the main one being, of course, the need they feel for an explanation of the lower average test scores of blacks, if, as the ETS claims, the tests are not culturally biased or unfair to minority students. Typically, escape-hatch explanations take the form of blame. Because the tests have been exculpated, blame is now redirected elsewhere, and in the ETS litany, the public schools, as usual, unfairly receive the brunt: "public miseducation," "failures of education," "teacher expectancy," and "caste and class barriers" to educational opportunity, along with "segregation" and "poverty." I have no way of knowing if this limited explanatory repertoire, which I have repeatedly encountered in the ETS's public statements, reflects an official doctrine of the ETS or merely a coincidental likeness of minds among those who speak for the enterprise.

Whatever faults one may legitimately point out in the public schools, the causation of the black IQ deficit certainly is not one of them. The typical one-standard-deviation mean difference between blacks and whites on tests of general intelligence or scholastic aptitude is full blown by the age of school entry, and it does not change

(relative to individual variability within the populations) from kindergarten to Grade 12. The schools, therefore, are not in any degree to blame for the observed social differences in scholastic aptitude. But should the schools be held culpable for not *overcoming* the difference? In the past 25 years, many millions of dollars of federal funds have been expended for numerous and massive attempts to overcome the difference, with apparently unimpressive success. The mean mental-test-score difference (in standard score units) between black and white youths is the same today as it was 75 years ago, at the time of World War I, when, for the first time, large samples of the nation's young men were given mental tests.

A standard individual or group test of general intelligence is an unbiased predictor of scholastic performance for blacks and whites, and it has proved no easier to raise intelligence and its correlated academic achievement by an appreciable amount for black children than for white children. We can safely say that, up to the present time, researchers have not yet discovered any educational prescription feasibility within the power of the schools that can substantially and permanently raise the general intelligence of black children *or* of white children. In this respect as in many others, the IQ difference between the races behaves very much as do the IQ differences among individuals of the same race. I have found no compelling evidence that the group differences are essentially different in nature from individual differences. The failure to discover any important race × treatment interactions (independent of IQ) in the educative process would seem consistent with this observation.

As for poverty and the other explanatory factors mentioned by Manning and Jackson, they should be viewed in the light of the available facts about the ETS tests: Less than 10% of the variance in SAT scores is *associated* with (not necessarily *caused* by) differences in family income; black students from the highest family-income level, on average, obtain SAT (and GRE, LSAT, MCAT) scores that fall at least half a standard deviation below the white average; and within any given level of measured aptitude, a higher percentage of blacks than of whites go to college.

Thus, these explanations in terms of popular clichés only lead eventually to embarrassment under critical scrutiny of the evidence. They are wholly out of keeping with the scientifically impeccable manner in which the ETS has treated the evidence pertaining directly to the tests themselves.

What would I suggest instead? Certainly, pusilanimous pussyfooting about the issue deserves no more to be condoned than pro-

pounding scientifically unfounded explanations. The simplest, most completely defensible course is the only scientifically honest one: *open agnosticism.* On this point I repeat what I said in *Bias:*

> The observed racial group differences are *real* in the sense that they are not merely an artifact of the measuring instruments. Once that point has been determined for any standard test . . . and the proper uses and limitations of the test are duly noted, the psychopsychometricians and the test publishers should be under no obligation to *explain* the causes of the statistical differences between groups. The problem of explaining the causes of group differences, aside from possible psychometric artifacts, is not the . . . responsibility of the constructors, publishers, or users of tests. The search for causes is an awesomely complex task calling for the collaborative endeavor of at least several specialized fields of science in addition to psychometrics. The state of our scientific knowledge on these matters at present only justifies an agnostic stance on the part of psychometricians, publishers, and users of tests, whatever else their personal sentiments may dictate. (p. 737)

MERCER

As I am told that Robert Gordon's chapter[7] is mainly addressed to an analysis of Mercer's position, I will only briefly indicate my views on a few key points of her paper.

Mercer's case, I believe, is built on what I have already referred to as the *sociologist's fallacy,* namely, the assumption of causality on the basis only of correlation. The whole notion of pluralistic population norms for tests of intelligence or scholastic aptitude is the full flowering of the sociologist's fallacy. Such norms are derived, essentially, by statistically adjusting the actually obtained test scores in terms of a number of their socioeconomic and cultural *correlates,* so that the derived scores for various subpopulations will be more nearly equal. The rationale for this procedure is based on the assumption that the subpopulations in question do not "truly" differ in whatever ability the test purports to assess, and that the observed differences in test scores merely reflect cultural differences. Minority groups obtain lower scores because of the "Anglocentric" bias of the tests. (Nothing is said about why Asians perform on a par with whites on these "Anglocentric" tests.)

If the usual standardized IQ test scores, instead of the pluralistically normed scores, are shown to be unbiased predictors of scholas-

[7]Editors' note: Gordon's chapter was written after Jensen's was completed; therefore, Jensen was unable to comment on Gordon's contribution.

tic achievement for majority and minority groups, then the derived scores from pluralistic norms are bound to be biased predictors. The preponderance of the present evidence indicates that the unadjusted IQs are unbiased predictors of scholastic achievement, whether measured by objective achievement tests or by teacher ratings. Where significant predictive bias has been found, it results in an *over*estimate of the actual performance of minority pupils (Messé, Crano, Messé, & Rice, 1979; Reschly & Sabers, 1979; Reynolds & Gutkin, 1980; Svanum & Bringle, 1982).

Mercer argues that the predictive validity of IQ for scholastic performance can be legitimately determined only from teacher ratings or graders, rather than from scores on achievement tests. The argument for this condition is that the correlation between IQ and achievement test scores is spuriously inflated by "common method" variance, because both measures are derived from tests. But it is hard to see that there could be much common method variance between an individually administered IQ test like the WISC or the Stanford–Binet and a paper-and-pencil scholastic achievement test. The lower correlation between IQ and teacher ratings of achievement than between IQ and scores on standardized achievement tests is explainable by (1) the lower reliability of teacher ratings and (2) the coarse scale, usually of only 3 to 5 points, on which teacher ratings or grades are assigned. This precludes as high a correlation as can be obtained between continuous variables measured on a fine-grained scale such as exists for IQ and standard achievement tests.

Under equal opportunity to learn, cognitive scholastic subject matter, after a course of instruction, will show individual differences that are highly correlated with scores on tests of general intelligence. This does not mean that a measure of scholastic achievement and an IQ measure of intelligence are one and the same thing. The contents and skills involved in the two tests may be "phenotypically" quite different. For example, proficiency in high school algebra is correlated with IQ, even though the IQ test items contain nothing resembling algebra and the IQ is measured before the pupils have taken a course in algebra or know anything at all about algebra.

The notion that we cannot make both a theoretical and a practical distinction between aptitude and achievement is nonsense. One of the more striking bits of evidence requiring such a distinction, which I have come across recently, is the finding by Carlson and Jensen (1981) of a correlation of $-.71$ between intraindividual (trial-to-trial) variability in choice reaction time (RT) and scores on one type of scholastic achievement: the Reading Comprehension test (Compre-

hensive Test of Basic Skills) among ninth-graders.[8] Where is the common method variance in this correlation? Or the common skills and knowledge? (Interestingly, the same RT measure was also correlated −.71 with scores on Raven's Progressive Matrices, a highly *g*-loaded nonverbal test.) A scientific explanation for such findings would not only justify but necessitate a distinction between ability and achievement. The fact that two classes of tests traditionally labeled *ability* (or *aptitude*) *tests,* on the one hand, and *achievement tests,* on the other, may in some cases be indistinguishable in appearance or, because of their high intercorrelation, can be used interchangeably for some purposes is beside the point. *Ability* and *achievement* are not different kinds of things, but different levels of analysis. The performances or achievements measured by all behavioral tests of whatever label are direct observations; abilities are inferred theoretical constructs needed to explain the observed covariation among a variety of performances and achievements.

"Edumetric" testing may supplant intelligence testing in schools, but I doubt that this substitution would make the controversy about bias in measuring general intelligence obsolete, as Mercer suggests. Instead, it will merely displace the controversy onto edumetric or scholastic achievement tests, because the largest part of the variance in these tests is identified by factor analysis as the *g* that is also measured by intelligence tests. The controversy over test bias will wane as educators and school psychologists gain greater understanding of the proper uses of intelligence tests and achievement tests and of the objective methods of assessing test bias. The largest study of item bias in scholastic achievement tests (Comprehensive Tests of Basic Skills), by Arneklev (1975), is reviewed in *Bias* (pp. 575–578). Out of 183 achievement test items were found a total of only 15 that met a statistical criterion of bias in large samples of black and white schoolchildren. Of these 15 biased items, 12 were biased in the direction that "disfavors" whites, and only 3 were biased in the direction that "favors" whites, in the effect of the item bias on the total score. Therefore, elimination of the biased items would slightly increase the average white–black difference.

In case anyone overlooks it, I should note the fact that the data in Mercer's Tables 6 and 9 can be used to examine Spearman's hypothesis of a correlation between tests' *g* loadings and the magni-

[8]An intriguing and researchable question is: Would the Reading Comprehension scores show the same or different regressions on the RT measures in white and minority groups?

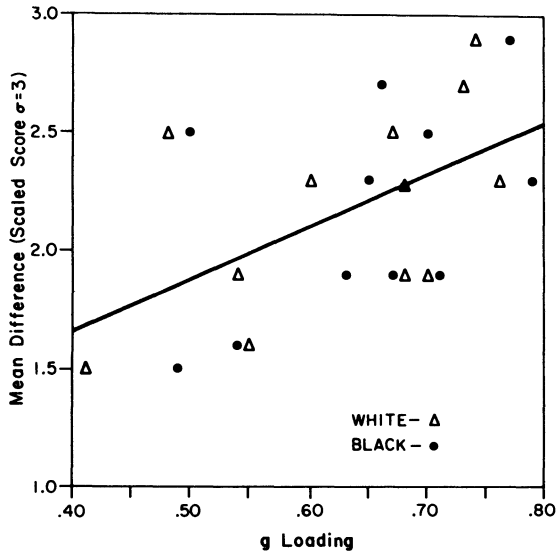

FIGURE 10. Mean white–black difference on 12 WISC-R subtest scores plotted as a function of the subtests' *g* loadings (corrected for attenuation) in Mercer's white (N = 683) and black (N = 638) samples.

tudes of the mean white–black difference on the various tests. The first principal component (Mercer's Table 6) is a good estimate of the WISC-R subtests' *g* loadings. These loadings are highly similar for blacks and whites, as indicated by a coefficient of congruence of .998 between the two sets of *g* loadings. For testing the Spearman hypothesis, these *g* loadings should be corrected for attenuation,[9] which I have done, using the reliabilities of the subscales based on the national standardization sample (N = 2,200). In Figure 10, the mean white–black differences (in scaled score units) on the WISC-R subtests are shown plotted as a function of the attenuation-corrected *g* loadings for whites and blacks. The correlation between the mean differences and average *g* loadings of blacks and whites is +.55, df = 11, $p < .05$. This correlation should be evaluated in light of the fact that in the national standardization data, the split-half sample reliability of the profile of white–black differences on the various subtests is only .83, and the corresponding split-half sample reliability of the *g* loadings is about .96. If these figures are used to correct the correla-

[9]This correction, however, makes little difference for these data, the correlations between the corrected and the uncorrected *g* loadings being .99 for whites and .97 for blacks.

tion of +.55 for attenuation, it becomes +.62. A significant positive correlation is consistent with the Spearman hypothesis, if the hypothesis is interpreted only as meaning that *g* is the most discriminating factor, racially. Similar data supporting the Spearman position that differences in *g* are primarily responsible for black–white differences on mental tests are reported by Reynolds and Gutkin (1981) and Jensen and Reynolds (1982). The obtained correlation of +.62 suggests that these two racial groups must also differ to some degree on other factors besides *g*.

EPILOGUE

The popular belief that all mental tests are necessarily culturally biased against racial minorities is well entrenched and of long standing. It remains to be seen how much longer this prevailing belief among nonspecialists in psychometrics will withstand contradiction by objective psychometric and statistical evidence and analysis. The words of Sir Francis Galton, generally considered the father of mental measurement and differential psychology, seem most appropriate here:

> General impressions are never to be trusted. Unfortunately when they are of long standing they become fixed rules of life and assume a prescriptive right not to be questioned. Consequently those who are not accustomed to original inquiry entertain a hatred and horror of statistics. They cannot endure the idea of submitting their sacred impressions to cold-blooded verification. But it is the triumph of scientific men to rise superior to such superstitions, to desire tests by which the value of beliefs may be ascertained, and to feel sufficiently masters of themselves to discard contemptuously whatever may be found untrue.[10]

REFERENCES

Ahern, S., & Beatty, J. Pupillary responses during information processing vary with Scholastic Aptitude Test scores. *Science*, 1979, *205*, 1289–1292.

Arneklev, B. L. *Data related to the question of bias in standardized testing.* Tacoma, Wash.: Office of Evaluation, Tacoma Public Schools, 1975.

Breland, H. M. *Population validity and college entrance measures* (Research Monograph No. 8). New York: The College Board, 1921.

Burt, C. *Mental and scholastic tests* (3rd ed.). London: P. S. King, 1921.

[10]This quotation appears on the frontispiece of every issue of the *Annals of Human Genetics*, which Galton founded in 1909.

Callaway, E. *Brain electrical potentials and individual psychological differences.* New York: Grune & Stratton, 1975.

Carlson, J. S., & Jensen, M. *Reaction time, movement time, and intelligence: A replication and extention.* Paper presented at the annual meeting of the American Educational Research Association, Los Angeles, Calif., April 1981.

Cleary, T. A., Humphreys, L. G., Kendrick, S. A., & Wesman, A. Educational uses of tests with disadvantaged students. *American Psychologist,* 1975, *30,* 15–41.

Elkana, Y. *The discovery of the conservation of energy.* Cambridge: Harvard University Press, 1974.

Equal Employment Opportunity Commission. Uniform guidelines on employment selection procedures. *Federal Register,* 1978, *43* (166).

Gordon, R. A. Labeling theory, mental retardation, and public policy: *Larry P.* and other developments since 1974. In W. R. Gove (Ed.), *The labeling of deviance: Evaluating a perspective* (2nd ed.). Beverly Hills, Calif.: Sage Publications, 1980.

Gordon, R. A., & Rudert, E. E. Bad news concerning IQ tests. *Sociology of Education,* 1979, *52,* 174–190.

Green, D. R., & Draper, J. F. *Exploratory studies of bias in achievement tests.* Presented at annual meeting of American Psychological Association, Honolulu, September 1972.

Hall, V. C., & Turner, R. R. Comparison of imitation and comprehension scores between two lower-class groups and the effects of two warm-up conditions on imitation of the same groups. *Child Development,* 1971, *42,* 1735–1750.

Hull, C. L. *Aptitude testing.* New York: World, 1928.

Humphreys, L. G. Intelligence testing: The importance of a difference should be evaluated independently of its causes. *Behavioral and Brain Sciences,* 1980, *3,* 347–348.(a)

Humphreys, L. G. Race and intelligence re-examined. *The Humanist,* 1980, *40,* 52–55.(b)

Jensen, A. R. *Educability and group differences.* New York: Harper & Row, 1973.

Jensen, A. R. Interaction of Level I and Level II abilities with race and socioeconomic status. *Journal of Educational Psychology,* 1974, *66,* 99–111.

Jensen, A. R. Race differences, strategy, training, and improper inference. *Journal of Educational Psychology,* 1976, *68,* 130–131.

Jensen, A. R. Genetic and behavioral effects of nonrandom mating. In R. T. Osborne, C. E. Noble, & N. Weyl (Eds.), *Human variation: Psychology of age, race, and sex.* New York: Academic Press, 1978.

Jensen, A. R. *Bias in mental testing.* New York: Free Press, 1980.(a)

Jensen, A. R. Chronometric analysis of mental ability. *Journal of Social and Biological Structures,* 1980, *3,* 103–122.(b)

Jensen, A. R. *Straight talk about mental tests.* New York: Free Press, 1981.

Jensen, A. R. The chronometry of intelligence. In R. J. Sternberg (Ed.), *Recent advances in research on intelligence.* Hillsdale, N.J.: Lawrence Erlbaum, 1982.(a)

Jensen, A. R. Reaction time and psychometric *g.* In H. J. Eysenck (Ed.), *A model for intelligence.* New York: Springer, 1982.(b)

Jensen, A. R., & Figueroa, R. A. Forward and backward digit span interaction with race and IQ: Predictions from Jensen's theory. *Journal of Educational Psychology,* 1975, *67,* 882–893.

Jensen, A. R., & Reynolds, C. R. Race, social class and ability patterns on the WISC-R. *Personality and Individual Differences,* 1982, *3,* 423–438.

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

Longstreth, L. E. Level I-Level II abilities as they affect performance of three races in the college classroom. *Journal of Educational Psychology,* 1978, *70,* 289–297.

Lorenz, K. *Die Ruchseite des Spiegels: Versuch einer Naturgeschichte menschlichen Erkennens.* Munich: R. Piper Verlag, 1973.

Matarazzo, J. D., & Wiens, A. N. Black Intelligence Test of Cultural Homogeneity and Wechsler Adult Intelligence Scale scores of black and white police applicants. *Journal of Applied Psychology,* 1977, *62,* 57–63.

McGurk, F. C. J. The culture hypothesis and psychological tests. In R. E. Kuttner (Ed.), *Race and modern science.* New York: Social Science Press, 1967.

Messé, L. A., Crano, W. D., Messé, S. R., & Rice, W. Evaluation of the predictive validity of tests of mental ability for classroom performance in elementary grades. *Journal of Educational Psychology,* 1979, *71,* 233–241.

Miles, T. R. Contributions to intelligence testing and the theory of intelligence: I. On defining intelligence. *British Journal of Educational Psychology,* 1957, *27,* 153–165.

Open peer commentary. *Behavioral and Brain Sciences,* 1980, *3,* 325–371.

Reschly, D. J., & Sabers, D. L. An examination of bias in predicting MAT scores from WISC-R scores for four ethnic-racial groups. *Journal of Educational Measurement,* 1979, *16,* 1–9.

Reynolds, C. R. The problem of bias in psychological assessment. In C. R. Reynolds and T. B. Gutkin (Eds.), *The handbook of school psychology.* New York: Wiley, 1982.

Reynolds, C. R., & Gutkin, T. B. A regression analysis of test bias on the WISC-R for Anglos and Chicanos referred to psychological services. *Journal of Abnormal Child Psychology,* 1980, *8,* 237–243.

Reynolds, C. R., & Gutkin, T. B. A multivariate comparison of the intellectual performance of blacks and whites matched on four demographic variables. *Personality and Individual Differences,* 1981, *2,* 175–180.

Reynolds, C. R., & Jensen, A. R. *Patterns of intellectual abilities among blacks and whites matched on g.* Paper presented at the annual meeting of the American Psychological Association, Montreal, 1980.

Reynolds, C. R., & Jensen, A. R. WISC-R subscale patterns of abilities of blacks and whites matched on Full Scale IQ. *Journal of Educational Psychology,* 1983, *75,* 207–214.

Schmid, J., & Leiman, J. M. The development of hierarchical factor solutions. *Psychometrika,* 1957, *22,* 53–61.

Schmidt, F. L., & Hunter, J. E. Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology,* 1977, *62,* 529–540.

Schmidt, F. L., & Hunter, J. E. Moderator research and the law of small numbers. *Personnel Psychology,* 1978, *31,* 215–232.

Schmidt, F. L., & Hunter, J. E. Employment testing. *American Psychologist,* 1981, *36,* 1128–1137.

Schmidt, F. L., Hunter, J. E., Pearlman, K., & Shane, G. S. Further tests of the Schmidt-Hunter Bayesian validity generalization procedure. *Personnel Psychology,* 1979, *32,* 257–281.

Silberman, C. E. *Crisis in black and white.* New York: Random House, 1964.

Spearman, C. "General intelligence," objectively determined and measured. *American Journal of Psychology,* 1904, *15,* 201–292.

Spearman, C. *The abilities of man.* New York: Macmillan, 1927.

Sternberg, R. J., Conway, B. E., Ketron, J. L., & Bernstein, M. People's conceptions of intelligence. Technical Report No. 28, October, 1980, and *Journal of Personality and Social Psychology: Attitudes and Scoial Cognition,* 1981, *41,* 37–55.

Svanum, S., & Bringle, R. G. Race, social class, and predictive bias: An evaluation using the UISC, WRAT, and Teacher Ratings. *Intelligence,* 1982, *6,* 275–286.

Van Valen, L. Brain size and intelligence in man. *American Journal of Physical Anthropology*, 1974, *40*, 417–423.

Vernon, P. A. Level I and Level II: A review. *Educational Psychologist*, 1981, *16*, 45–64.

Wherry, R. J. Hierarchical factor solutions without rotation. *Psychometrika*, 1959, *24*, 45–51.