THE BEHAVIORAL AND BRAIN SCIENCES (1980) 3, 325–371
Printed in the United States of America

Précis of *Bias in Mental Testing*

Arthur R. Jensen
Institute of Human Learning, University of California, Berkeley, Calif. 94720

Abstract: Most standard tests of intelligence and scholastic aptitude measure a general factor of cognitive ability that is common to all such tests – as well as to all complex tasks involving abstraction, reasoning, and problem-solving.

The central question addressed by this inquiry is whether such tests are culturally biased in their discrimination between majority and minority groups in the United States with respect to the traditional uses of such tests in schools, college admissions, and personnel selection in industry and the armed forces.

The fact that such tests discriminate statistically between various subpopulations does not itself indicate test bias. Acceptable criteria of bias are based on (1) the test's validity for predicting the performance (in school, on the job, and so on) of individuals from majority and minority groups, and (2) the internal consistency of the test with respect to relative item difficulty, factorial composition, and internal consistency/reliability.

A review of empirical studies relevant to these two criteria reveals that the preponderance of evidence contradicts the popular belief that the standard tests most widely used at present are culturally biased against minorities. The tests have the same predictive validity for the practical uses of tests in all American-born, English-speaking racial and social groups in the United States.

Factors in the test situation, such as the subject's "test-wiseness" and the race of the tester, are found to be negligible sources of racial group differences.

Keywords: cultural bias; intelligence; IQ; mental tests; race differences; sex differences; test bias

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 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; (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* factor), and he developed a mathematical technique, known as *factor analysis*, that made it possible to determine

 0140-525X/80/030325-47/$04.00/0

(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., percent 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 layman thinks of as "intelligence?" Yes, very largely. Persons whom laymen 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 sizeable 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 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 percent of the white as compared with less than 3 percent 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 occupa-

tions 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, were 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 involves 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 or not 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 level of developed ability for doing college work, because it is the student's developed ability that actually predicts his future success in college, and not some hypothetical estimate of what his ability *might* have been if he 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 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 supported.

Misconceptions of test bias

There are three popular misconceptions or fallacies of test bias which 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 holds that any test which 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 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-cent 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 or not 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! For example, the Verbal Comprehension subtest of the Wechsler IQ scales is frequently singled out as an example of a culturally unfair set of items. Yet blacks score higher on the Verbal Comprehension test than on any of the other eleven subscales of the Wechsler, with the exception of the digit span memory test. 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 (percent passing) in the white population (McGurk, 1967).

Standardization fallacy. This is the belief that a test which 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 which is expressed in terms of the coefficient of *reliability* (i.e., the proportion of the total variance *not* attributable to random errors 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 which have a different meaning for members of one group than 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 possible indicators of bias that a test "passes" without statistical 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, since 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. These 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 corrected – 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

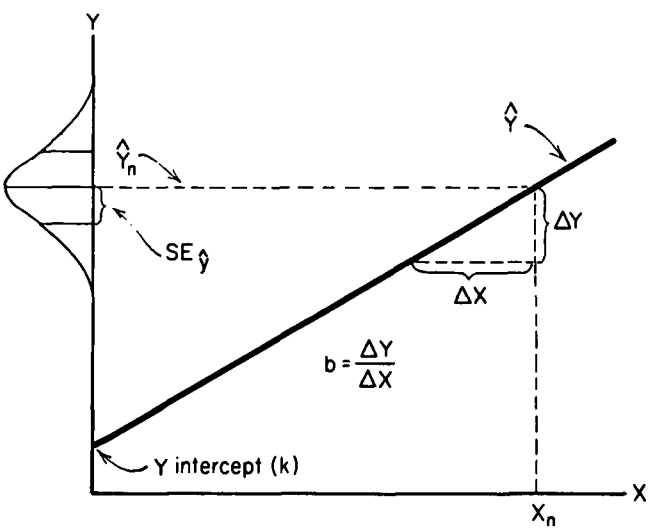

Figure 1. Graphical representation of the regression of criterion measurements (Y) on test scores (X), showing the slope (b) of the regression line $\hat{Y}$, the Y intercept (k), and the standard error of estimate ($SE_{\hat{Y}}$). An individual with a test score X_n would have a predicted criterion performance of $\hat{Y}_n$, with a standard error of $SE_{\hat{Y}_n}$. 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 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.

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

Internal indicators. These are psychometric features of the test data themselves, such as the test's internal consistency reliability (a function of the inter-item correlations), the factorial structure of the test or a battery of subtests (as shown by factor analysis), the rank order of item difficulties (percent 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 these essential psychometric features of tests should not differ between populations for an unbiased test.

Undetectable bias. Theoretically there is a type of bias which 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 which 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.

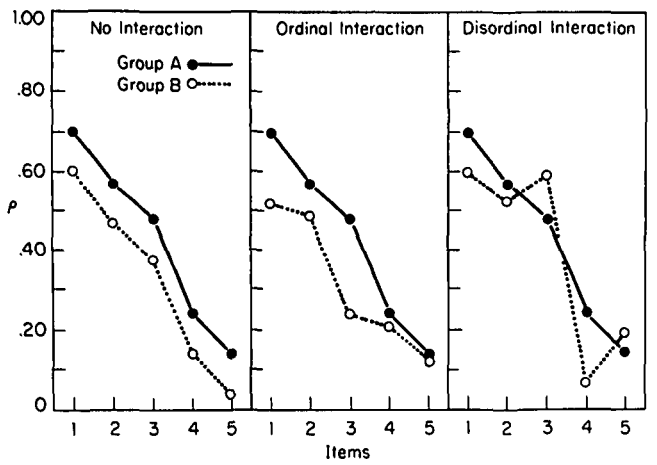

Figure 2. Graphic representation of two types of items × groups interaction for an imaginary 5-item test. Item difficulty (proportion passing the item) is shown on the ordinate, the 5 items on the baseline. When the item difficulties for two groups, A and B, are perfectly parallel, there is no interaction. In ordinal interaction, the item difficulties of groups A and B are perfectly 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.

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, 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 which 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 shall 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 which is derived on an all-white or predominately 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 – 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 eliminated 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 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, & Wesman, 1975; Linn, 1973).

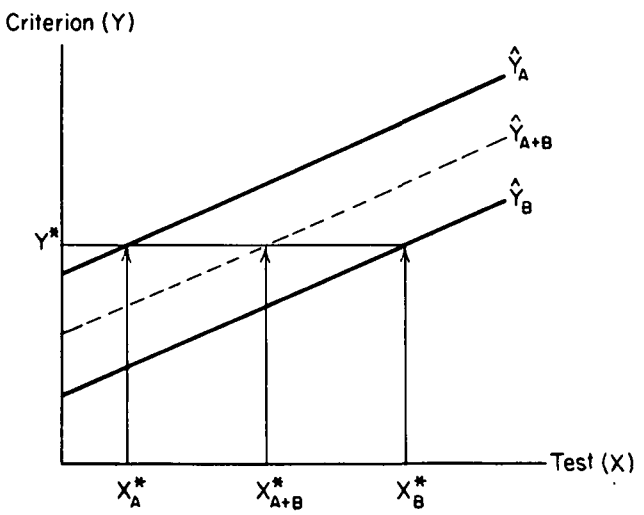

Figure 3. An example of the most common type of predictive bias, viz. intercept bias. The major and minor groups (A and B, respectively) actually have significantly different regression lines Y_A and Y_B; they differ in intercepts but not in slope. Thus, equally accurate predictions of Y can be made 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 underpredicted, and the performance of individuals in group B (the lower-scoring group) will be overpredicted, i.e., 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.

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 eleven 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 is FDS. There is no plausible *cultural* explanation for this phenomenon.

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.

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 (percent 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 complexity, 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 than 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 than 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 between blacks and whites, and is seldom less similar than the similarity between two random halves of either the white or 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), 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: (*a*) correlation of the item with age (younger versus older children); (*b*) correlation of the item with ability in children of the same age as determined by total score on the test; and (*c*) correlation of the item with race (white versus black). We then obtain the correlations among *a*, *b*, and *c* 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: (1) 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; (2) in both groups, the items that correlate the most with age are the same ones that correlate the most with ability; and (3) 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 they were 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 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 it 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 rather 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 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). [See Brainerd: "The Stage Question in Cognitive–Developmental Theory" BBS 1(2) 1979.] 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 the 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 dimension selected for matching figures on the basis of color, shape, size, and number, five- to six-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, and 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, suggest 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 which 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 to 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 thirty 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 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, 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 [see Rosenthal & Rubin: "Interpersonal Expectancy Effects" *BBS* 1(3) 1978].

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

Conclusion

Good tests of abilities surely do not measure human worth in any absolute sense. But they do provide indices which 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 American-born, English-speaking segments of American society today, regardless of sex or 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 which are causally independent of them. 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, 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 specialists in several 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 application of sound psychometrics can help to reinforce the democratic ideal of treating every person according to his or her *individual* characteristics, rather than according to sex, race, social class, religion, or national origin.

Open Peer Commentary

Commentaries submitted by the qualified professional readership of this journal will be considered for publication in a later issue as Continuing Commentary on this article.

Editorial Note. *The next issue of BBS will include a treatment of a related topic: R. Sternberg's "Sketch of a Componential Subtheory of Human Intelligence"* BBS *3(4) 1980.*

***by* C. Loring Brace**

Museum of Anthropology, University Museums Building, Ann Arbor, Mich. 48109

Social bias in mental testing

Arthur R. Jensen has assured an interest in his work by his previous association with the view that the average difference between the IQ scores of black and white Americans is largely genetically determined and that, on the whole, blacks are innately less intelligent than whites. Whatever his original intentions, the result has been a searing controversy that has continued unresolved. Over the last ten years, his work has been subjected to more than the usual amount of critical scrutiny. One of the consequences is that a great many more people will read this book than would normally be expected to plow through a manual on the statistical properties of the professional tester's armamentarium.

Clearly this tome represents the product of a prodigious amount of effort. Furthermore, its production required no small amount of courage. Jensen has been the subject of scathing criticism, and, although I feel that it has been thoroughly deserved, this means that anything that he writes for the public domain is sure to be examined from the perspective of a remembrance of his previous critizied efforts. Cognizant of all of this, Jensen has prepared his latest opus with meticulous care. It is impeccably edited and clearly written. This in itself is no minor achievement in a subject that is so relentlessly statistical. Terms are clearly defined and procedures are explained in simple and straightforward fashion, and the author has resisted the temptation to hide his justifications in a cloud of mathematical legerdemain.

In spite of all of this, however, and in spite of some laudable if belated statements concerning democratic ideals of fariness and the rights of people to be treated as individuals, this work can be viewed as a monument to the skills of artful dodging. At the very beginning, Jensen attempts to defuse potential criticism by noting that psychometric tests measure phenotypes and not genotypes, and that the *Bias in Mental Testing* referred to in the book's title concerns only the question of bias in the measurement of phenotypes. He goes on to declare, "We need not be concerned with inferred genotypes in this inquiry" (p. xi). This is repeated in the brief but well-considered section on the concept of heritability and its attendant problems (pp. 243–245). One would never guess from this that the reason Jensen is regarded so critically by many scientists is precisely because of his earlier work, in which it was felt that he disastrously misused the concept of heritability. That earlier paper is missing from the bibliography and no mention is made of the nature of the criticism that it elicited from geneticists, psychologists, and other social scientists.

Then, after hundreds of pages of what is presented as scrupulously fair and straightforward treatment, he concludes by assuming what was initially denied to be the focus of concern. The measured difference in "ability" between blacks and whites is too great, he would have us believe, to be accounted for by differences in schooling, language, or familial and cultural background. The difference remains, however, and it is taken to indicate that blacks and whites simply come equipped with different amounts of *g* or "intelligence."

If this is one example of artful dodging, there are many others. Some of these are by omission, which is something that the reader who is new to the technical literature on the subject would not suspect from the appearance of massive documentation with which the book is presented. Black-white differences are presented as eternal verities with no depiction of context. No mention is made of the fact that there is an urban-rural spectrum of white difference which is quite comparable to the black-white difference that is the main focus. No mention is made of the differences between northern and southern white scores, northern and southern black scores, World War I versus World War II white armed forces averages, or identical twins raised apart in families of different socioeconomic status. Nor is the fact mentioned that the blacks on whom most of the conclusions were established were largely poor and southern. Nor is any mention made of the fact that precisely the same procedure and "logic" was used in the 1920s to establish the innate intellectual inferiority of "non-Aryan" Europeans. The consequences ranged from the immigration restriction legislation of 1924 to the lingering and tasteless legacy of "Polish jokes" and other derogatory ethnic characterizations.

From the anthropoligical point of view, one can grant that Jensen's technical approach to the nature of test bias may allow him to note that this is not a factor in the creation of "race" differences, but it does not absolve him from questions concerning the bias of the tester who creates the test dimensions in the first place. The nature of Jensen's

own bias is clearly shown in chapter 9, especially on pages 370–372, where he discusses what he calls "Inadequate Concepts of Test Bias." Here he enumerates three points which by fiat he designates as fallacies. The third one, his "standardization fallacy," is properly treated and need not concern us further, but the other two require more comment.

The first of these, the "egalitarian fallacy," is the "assumption of equal or equivalent intelligence across all human populations," which he simply declares to be "gratuitous" and "scientifically unwarranted." In his words, "The egalitarian assumption obviously begs the question in such a way as to completely remove itself from the possibility of scientific investigation" (p. 370).

First of all, the assumption of equal ability is neither gratuitous nor scientifically unwarranted. The student of human evolution observes that a hunting and gathering way of life was the common human heritage for more than a million years, only being abandoned within the last 10,000 years. The shaping effects of this should have given all modern human populations essentially equivalent mental capacities. For similar reasons, ethologists who study wild canids have no reason to postulate differences in intellectual capabilities that would distinguish the Norwegian from the Alaskan wolf, or either from the Australian dingo.

As for the matter of scientific investigation, over a decade ago the Society for the Psychological Study of Social Issues noted that this would "be possible only when social conditions for all races are equal and this situation has existed for several generations" (May 2, 1969). Jensen's subsequent reply was that, "Since no operationally testable meaning is given to 'equal' social conditions, such a statement, if taken seriously, would completely preclude the possibility of researching this important question, not just for several generations, but indefinitely" (Jensen 1971, p. 24). As his present book demonstrates, he has simply continued in his efforts to prove the existence of the differences he assumes, with a minimum of attention to whether or not the social conditions are comparable.

The second of his "fallacies" in chapter 9 is his "culture bound fallacy." This he restricts to the possibility that certain test items will be less familiar to members of one culture (or subculture) than to those of another. It is a legitimate concern, but what is completely lacking is a regard for the possibility that a perception of the world as an itemized construct may differ markedly from one culture to another. If this is true – and there is evidence to suggest that such is the case – then no amount of item substitution or manipulation can make the test into an equivalent investigation of the inherent capacities of the groups being studied.

It seems appropriate here for me to repeat the conclusion which I voiced when this matter first arose: "if in fact Jensen were really interested in an unbiased testing of the heritable component of intellectual differences between human groups, he should have been devoting his efforts to setting up a scientifically acceptable test situation. The very first step would involve engaging in an attempt to produce an operational definition of equal social conditions and the systematic effort to see that these be extended to all of those whom he might wish to test." "Then, and only then, could the question of inherited differences in ability be posed. In fact, whether or not the question is indeed 'important' could only be decided under such circumstances" (Brace 1971, p. 8).

If the view of the world through the rigid categories of the literate upper middle class in western society seems uncomfortably narrow and self-satisfied to some, it is clear that certain of Jensen's intentions are good. He articulates one theme at the very beginning which justifies our continuing attention whether or not his book can be regarded as an adequate treatment. "Before the use of tests is rejected outright, however, one must consider the alternatives to testing – whether decisions based on less objective means of evaluation (usually educational credentials, letters of recommendation, interviews, and biographical inventories) would guarantee less bias and greater fairness for minorities than would result from the use of tests" (p. ix). This is an issue that will not go away, and if we are not satisfied with Jensen's treatment, it is just further proof of the fact that it is something that is too important to be left in the hands of the experts.

If we mull over the totality of Jensen's work, there is a fundamental aspect of it that sticks in the mind of the reader with a sense of intellectual history. Dating from the time of the medieval logicians and reaching its apotheosis in the *magnum opus* of Archdeacon William Paley more than a century and a half ago, there was a current of thought which assumed that the regularities visible in the organic world constituted evidence for the existence of God. From the manifestations of design, one could infer the existence of a designer. Now, in the works of Jensen and others, we seem to have a psychometric reincarnation of that idea. Although statisticians repeatedly warn us that we cannot make the leap from correlation to cause, it would seem that psychometricians have been beguiled by indices of covariance to infer the existence of a single but not directly ascertainable entity labelled intelligence. In this case, however, whether Spearman intended it this way or not, their divinity is depicted with a lower-case *g*.

In the summary of chapter 6, Jensen declares that, "A working definition of intelligence . . . is that it is the *g* factor of an indefinitely large and varied battery of mental tests." The concluding sentence of that chapter, however, states that, "At present, it seems safe to say, we do not have a true theory of *g* or intelligence." Despite this, there is no weakening in the faith that such an entity exists and that various human groups possess inherently different amounts of it.

If this manifestation of unprovable belief owes less to the reverence for holy writ than did the natural theology of Archdeacon Paley, it owes less also to the "nature" from which Paley sought his proof. Instead, the wellsprings of faith derive from the proliferating realm of mental tests accompanied by a litany of "construct validity," "criterion variable," "delta decrements," and much more. Problems that cannot be treated psychometrically are clearly regarded as being of lesser importance and are summarily dismissed to the realm of the moral, legal, or philosophic. The view which Jensen articulates with painstaking thoroughness and the enthusiasm of the devout could well be labelled statistical theology.

With all the zeal of the true believer, Jensen has compiled an exhaustive treatment of the realm in which he has committed his faith. It is a technical realm, and within it he is a consummate technocrat. The key to his approach is made clear in the preface when he says, "One cannot treat a fever by throwing away the thermometer" (p. xi). His answer, as one might expect, is unrelated to finding out about the causes and treatment of fever, but in the devising of more ingenious and "unbiased" thermometers. Curiously, he does not continue with his metaphor. The fever thermometer measures the response of the body to the insults of the environment, and not even the most devout believer in the inevitability of human racial differences has questioned the egalitarian assumption where normal human body temperature is concerned. Is it not just possible that the ingenuity of the psychometricians, in comparable fashion, has done no more than demonstrate the effects of different conditions on what, in a populational sense, is a common heritage of human mental capacity?

This is not just a trivial and obfuscating suggestion. There are many more reasons for choosing it as a starting point for investigation than for choosing the converse. Yet, for those who have assumed that mental reality exists in test scores standardized on middle-class American whites, such a possibility has never even been mentioned, let alone given serious consideration. To modify that famous concluding sentence of a century ago to suit the present limited perspective: "There is myopia in this view of life" (Darwin, 1859, p. 490).

by **Hunter M. Breland**

Educational Testing Service, Princeton, N.J. 08541

Population validity and admissions decisions

The author has chosen to tackle the issue of test bias head on, as the title of his book suggests. In my own work I have taken a different approach. The term "test bias" is an unfortunate one, both because definitions of it vary and because it is not always clear which of the existing definitions, or what new definition, is being used. To tie the word to tests is also confusing because it implies that tests are used in isolation to make important decisions about individuals, such as

admission to an institution of higher education, and that these decisions are therefore biased. Less complicated are discussions of validity as they relate to the various populations of interest and decisions about them. Validity is concerned with the degree to which accurate inferences are made from data, and thus population validity is concerned with the accuracy of inferences about individuals who are members of identifiable populations. A decision to admit or not to admit an individual to an institution of higher education is based upon inferences made from data. I restrict my commentary to those parts of the book pertaining to population validity and admission to institutions of higher education.

My first comment is with respect to the assertion that "the tests screen out a larger proportion of black than of white applicants" (p. 482). It is the belief that this statement is true that leads to a generally defensive posture in the exposition. In fact, much evidence would suggest that the statement does not hold, in recent years at least. Were tests used solely as the basis for admission, if they were used in precisely the same way for all populations, and if future performance of individuals were the only criterion of judgment, then it would probably be the case that the tests would exclude proportionately greater numbers of some populations. It is perhaps because none of these conditions holds that the evidence does not support the statement. Only a narrow view of validity and admissions, as in the traditional test bias approach, would lead to such a conclusion.

It is a similarly narrow view that occasions my second comment. The author suggests (p. 469) that the Scholastic Aptitude Test (SAT), the Medical College Admission Test (MCAT), the Law School Admission Test (LSAT), and the Graduate Record Exam (GRE) all assess essentially the same thing because scores on them are correlated. At another point (p. 317) the same argument is refuted by the author's own words: "diameters and circumferences of circles are perfectly correlated, but no one would claim that they are the same thing." The notion of validity requires far more than correlational and empirical evidence. If this were not so, one could devise admissions procedures based solely on a single intelligence test. There would be no need for committees of experts to make judgments about and to contribute to the formulation of procedures for admission to their profession.

It could not be expected that a book so broad in scope could hope to capture all of the relevant literature within each of the many areas covered. Nevertheless, it is useful to note some important studies that were missed. Borgen (1972) illustrates one of the important pitfalls in validity research. This is the problem encountered when widely varying college grading standards are pooled together to generate so-called "validity coefficients." The author of the book does not note this problem or cite references pertaining to it, but it is a rather crucial point of information for those attempting to interpret validity. The point is that if one pools too much data the resulting correlation coefficients often become meaningless. Despite this omission of a 1972 paper, the author does remarkably well up to about 1976, where his review tends to end. There are a number of papers of about this era that were missed: for example, Goldman and Widawski (1976), Goldman and Hewitt (1976), Farver, Sedlacek, and Brooks (1975), and Silverman, Barton, and Lyon (1976). Of course, other relevant papers have appeared since 1976, but it cannot be expected that a 1980 book can encompass all that came before it in time.

There are also those obscure papers and reports that no work can hope to include unless it is very precisely focused and unless the author has unusual contacts with those engaged in the field. In this connection, it is useful to note that in one area of interest, that of the objective criterion, there has been some work that the author is not aware of. On page 488 the author notes that "only one white and black college prediction study has used a more exact criterion than GPA" (grade point average). The study noted is the Centra, Linn, and Parry (1970) examination of SAT-V and SAT-M as predictors of the Area Tests (Achievement Tests) of the GRE. In addition to this kind of objective criterion study, there have been studies of the prediction of essay-writing performance across groups (Breland 1977), in which a multiple choice test was shown to predict equally well both minority and nonminority performance in college in writing brief essays. Further elaboration of these and other studies is given in Breland (1979).

One final comment is that the author (or his research assistants) exhibits some ignorance when a study by Munday (1965) on "the American Council on Education Test (ACE)" is reported (p. 485). What is intended is the American College Testing program's ACT test.

Despite the critical comments made here, I suspect that the book is one that will receive a great deal of study over the years to come. Both professors and students will find it a useful resource for the resolution of a myriad of arcane issues relating to testing.

by **Nathan Brody and Erness B. Brody**

Department of Psychology, Wesleyan University, Middletown, Conn. 06457

Differential construct validity

Jensen's book presents an exhaustive and definitive demonstration that black-white differences in scores on intelligence tests are not a function of item type. The difference exists for virtually any items used on any of the standard tests of intelligence. Rather than dwell on this central point of agreement, we would like to mention one specific point of disagreement and then indicate where we believe that a somewhat different emphasis is justified.

Jensen presents data which he believes support the "Spearman hypothesis," which asserts that racial differences will be largest on those tests that are pure measures of *g*. As he notes (Jensen 1980, pp. 548–549), the available data based on analyses of factor loadings in batteries of tests fail to discriminate between a loading as a statistical artifact of the composition of a battery of tests and a conceptual definition of *g* in Spearman's sense of the "education of correlates and relations." His discussion of evidence with respect to the second, and more central, aspect of the Spearman hypothesis relies on the use of tests which are only tangentially relevant to Spearman's conceptual definition (e.g., forward versus backward memory span, simple versus complex reaction time, multifactor batteries combining tests of achievement and ability) and omits data which provide a direct test of the hypothesis. Jensen asserts that the Raven test is a good measure of *g* and the Peabody Picture Vocabulary Test (PPVT) is not only subject to cultural bias but is also a poor measure of *g* (see Jensen 1974a). His own data, involving a large sample, indicate white-black differences expressed in standard deviation units of .88 on the PPVT and .86 on the Raven (Jensen 1974a). These data provide a simple and direct contradiction of the "Spearman hypothesis."

Jensen's discussion of construct validity relies excessively on factor analytic studies, item analyses, and predictive validities. While it is true that the construct, "intelligence," is in part definable in these terms, it is also true that the construct has a conceptual basis. For example, a definition of intelligence as ability to learn implies not only a predictive relation between intelligence test scores and academic achievement but a causal relation as well. Jensen cites the study of Crano, Kenny, and Campbell (1972) which uses cross-lagged panel analysis to demonstrate support for the expected causal linkage underlying the construct validity of intelligence tests. However, Crano, Kenny, and Campbell found supporting evidence for construct validity only in their suburban, predominantly white, sample. In their predominantly black sample there was no evidence of construct validity, implying that intelligence tests may not be adequate measures of the construct 'intelligence' in black samples. Similarly, the term 'ability' is often taken to imply a characteristic which is at least in part heritable. We know relatively little about the possibility of differential heritability of scores in different populations. Fischbein (1980) has provided data for a Swedish sample indicating that the heritability of test scores may be lower in disadvantaged samples than in advantaged samples. He finds, for example, that the intra-class correlations for monozygotic and dizygotic twins in his most advantaged social group on a test of verbal ability are .755 and .374 respectively. The comparable correlations are .661 and .509 for his least advantaged group. These differences suggest lower heritability for intelligence in less advantaged groups, and, if confirmed with respect to black-white differences, would also provide data relevant to the possibility of differential construct validity of test scores in different groups.

It is possible to argue that Jensen exaggerates the social and

practical relevance of the racial difference in intelligence test scores. Although the test scores predict school grades, this prediction may in fact be a function of the structure of schools rather than an intrinsic or necessary relationship. Bloom (1976) has shown that the correlation between intelligence test scores and school learning can be reduced under mastery learning conditions. Moreover, the correlations between IQ and occupational performance in most jobs are relatively low. For example, Jensen's discussion of the ETS-Civil Service Commission study (Jensen 1980, pp. 505–507) indicates that the predictive validities of ability tests for work sample composites rarely exceed .30. Correlations in the 1930s (see also Jensen's discussion of Ghiselli's monograph and the Hum RRO study, Jensen 1980, pp. 348–351) indicate that the percentage of variance accounted for by tests even for relatively skilled occupations is not high. Moreover, as Bloom's work has indicated, the relation between test score and a criterion is a variable rather than a fixed entity. Thus, changes in training procedures or in what may be relatively unimportant characteristics of jobs may serve to reduce the importance of intelligence as a determinant of job success. (For a further discussion of limitations in the use of IQ tests, see Brody & Brody 1976, ch. 7.)

by **Raymond B. Cattell**

Department of Psychology, University of Hawaii, Honolulu, Hawaii 96822

"They talk of some strict testing of us – Pish"

In the last few years Dr. Jensen has given us three books outstanding for their scholarship and constructive reasoning: *Genetics and Education* (1972), *Educability and Group Differences* (1973), and *Bias in Mental Testing* (1980). The last – with which we are here concerned – completes a trilogy of interdependent works likely to be considered a landmark and a turning point in educational theory and practice.

The current level of education on behavior genetics being what it is, no observer can be surprised at the ill-informed vehemence with which Jensen has been attacked. Most of these attacks he has answered, in closely reasoned articles. But it was inevitable that the flood should ultimately fling itself against the very data base of most inferences, namely, mental testing and the concepts of psychometry.

The psychometrics of personality and ability stands as the most scientifically developed pillar of psychology (with perhaps learning theory as the second). This has not prevented its coming under greater public attack in the last two decades than the flimsier clinical elaborations of psychology from which the public more seriously suffers. The causes of the hostility to mental testing deserve a thorough diagnosis by an able social psychologist, but the present writer would contingently point to three: (1) A failure of psychologists themselves to teach an adequate level of sophistication about tests, personality theory, and ability theory to those engaged in testing. (2) The rise in the sixties of what has been called the "me generation," living on the pleasure principle, and declining to meet objective standards. Like the drunkard in Omar Khayyam they exclaim, "They talk of some strict testing of us – Pish!" One had to argue with these abolishers of all examinations: "Would you like to be operated on by a dentist or doctor from a school with no examinations?" (3) It has been claimed that since some minorities perform more poorly there must be test bias. "Minority" gets absurd definitions, as when women are called a minority. Women perform equally on intelligence tests, below males in spatial ability but above in Thurstone's verbal factor. Several minorities, for instance, Chinese and Jews, perform above the American majority. The fact of belonging to a minority culture does not in itself produce poorer test performance.

It is surprising that when examinations and testing came under fire from these diverse bands of metrophobic guerrillas the main constructors and users of tests did not explain to the public what tests do. They apparently felt themselves secure enough to reply by a dignified silence. It has fallen to Jensen to meet the issues, and he has done a thorough job from A to Z, meeting the theoretical psychometric and practical social problems, and to some extent even the philosophical background issues.

After recognition of the criticisms, Jensen devotes three chapters to the varieties of abilities and their distribution, centering largely on the definition of intelligence. He proceeds to technical psychometric discussion of reliability and validity and thereafter to five chapters intensively examining the areas and forms of possible bias. He concludes with discussion of remedies in the form of culture-reduced tests, and so on.

In none of these fields is it possible to fault Jensen for any lack of scholarship and acquaintance with an up-to-date view of technical matters. I shall mention a few places, nevertheless, where my own emphases would be slightly different. Jensen clearly recognizes that there are two distinct problems in comparing the performance of two groups: (1) Is the factor structure of the abilities the same in the two populations, and (2) Are the mean levels and distributions the same?

Jensen recognizes the fundamental change in intelligence theory over the past twenty years, from a single general factor, Spearman's g, to two distinct general factors, fluid intelligence, g_f, and crystallized intelligence, g_c. The first is more brain-growth-dependent and has higher heritability. The second pattern is the result of the investment of the first in the learning of complex material, for example, grammatical skill, social situation rules, mathematics. A whole series of different properties – in age curves, IQ sigma, relation to brain damage, and so on – distinguish these two g's. However, they correlate, 0.3 to 0.6, depending on age, social background, and so forth, the positive correlation indicating that it is the investment of g_f in learning experiences that begets g_c.

The traditional intelligence test measures g_c. It will obviously vary in its loadings on any subtest according to the subculture. To show the futility of trying to determine individual endowment differences by such a test as the WAIS and the WISC in different subcultures, one has only to think of giving such a test to a sample of more mixed cultures – say some Americans, French, Pakistani, and Russian. The calculated heritabilities, for example, would be widely different. By contrast, culture-fair scales have subtests – classifications, analogies, series, and so on – all in fundaments of immediate perceptual identity – that yield the same loading pattern across cultures, as a comparison of, say, Weiss's (1972) factoring in Germany and those in America, or Horn and Cattell's (1978) comparison of factor patterns for white and black children, shows. A sufficiently complete clarification of issues is possible only with such tests of fluid intelligence, where the psychologist knows that he is measuring the same thing, and where verbal, mechanical, numerical, and other acquired skills do not enter to create false estimations of level. On the other hand, within America, as Jensen shows, there is no evidence of really significant differences of loading pattern for g_c among different minority groups. And, as Humphreys (1973) and others have shown, the regression of such measures on school performance, that is, the predictive value of intelligence tests for school achievement, is indistinguishable for white and black.

What *is* different is the absolute level of test performance – by an amount of approximately 15 points of IQ, which has apparently remained unchanged over many years, since Shuey (1966) first began collecting data. Stable though this difference is, it naturally remains intrinsically suspect, because g_c has so substantial an amount of acquired cognitive skill in it. As Jensen indicates, however, the application of culture-fair tests, though recent (see, for example, Knapp 1960), points to differences in the same direction as with traditional tests, and in some cases quite as large. Indeed, Jensen points out that on verbal tests blacks come closer to the white average than on culture-fair measures.

This last is not surprising, for it has long been recognized that on measures of Thurstone's primary abilities different ethnic and racial groups have different profiles. For example, Chinese are notably higher on numerical and spatial skills than on verbal and Jewish children exceed the American average most on verbal skills (Cattell 1971, p. 292). Not enough is yet known about the developmental role of general intelligence in these primary abilities, or about action from genetic or environmental sources, to decide how far they are genetic, but Stafford (1961) has made an argument for spatial ability having appreciable heritability, and being sex-linked in its chromosome location.

As far as Jensen's position on ability structure and ability test validities is concerned, the present reviewer can disagree only over

trifles. One of these would be the acceptance (in which Jensen is not alone) of measures on the Raven matrices test as good measures of fluid intelligence. It is clear that the different subtests in a culture-fair test *do* have specifics in them, over and above g_f, but what they are we do not yet know. A good culture-fair intelligence test would therefore employ several (say, four or five) different subtests, for example, series, analogies, matrices, and classification, to wash out any undue contamination by one specific one.

Such trivia aside, Jensen's chapters 5 and 6 provide as incisive, balanced, and complete a study of intelligence and intelligence tests as the student is likely to find anywhere in so few pages. These chapters alone would make the book worthwhile, even indispensable, for the student who would be up to date.

The immediately succeeding chapters 7, 8, and 9 constitute an equally penetrating, complete, and brilliantly clear account of the psychometric and statistical aspects of ability testing – indeed of most testing. Here we have a subbook of three chapters that will be very helpful to psychology students even when they are not concerned with the main theme of the book. The present reviewer, who has for some years been trying to establish a clarity of concepts in the test *consistency* field, distinguishing the reliability (or *dependability*) coefficient from the *homogeneity* and *stability* coefficients (particularly in evaluating emotional state measures), finds the treatment of these issues, and especially of the stability coefficient calculation and meaning, delightfully clear.

Although behavior genetics per se is not an intrinsic part of a book on bias in mental testing, certain inferences in the latter area require reference to the former. Here one encounters a virtuosity of teaching performance arising naturally from Jensen's long experience in that area. Essentials are stated clearly and with apt illustration – as well as with the correction of some misuses of formulae that have been rather prevalent. Since these references are scattered one cannot refer the student to a given chapter, as one can for a good condensed overview of intelligence and of test psychometrics, but he will absorb some central behavior genetics "through the pores" as he reads on these other themes.

On the definitions of *criteria* and evaluations of *test bias*, chapters 10 through 13 are very comprehensive. Whereas the preceding chapters are mainly technically clear and penetrating accounts of concepts not essentially new, the definition and ordering of concepts of bias in these chapters constitute a very creative contribution to test theory, building a new wing onto that structure. The only addition the present reviewer would like to have seen is an examination of how far the model of *trait view theory* might comprehensively handle certain classes of bias. Trait view theory treats that part of the error which arises from the different situations (of motivation, fatigue, interest) in which the actual testing is done. Fulker (1973), for example, in connection with his genetic studies of intelligence, has asked what error or bias may arise from particular conditions in actual *testing situations*.

Finally we come to Jensen's handling of the social and ethical problems, which is somewhat more extended and direct here than in his earlier writings. Previously he has first stood by the duty of a scientist to publish results as he finds them, and secondly has proposed ameliorative steps for the educational problems involved. As to the first, some journalists have presented him to the public as a person tendentiously involved in the problem of minorities, whereas those who have long known Jensen recognize that he stumbled over these issues in the course of scholarly research. As to the second, he has proposed several educational practises to reduce the conflicts not only of minorities but of individuals below average intelligence. His separation of A and B kinds of school curriculum performance – those requiring much and those requiring less intelligence – and the recognition that all can hope to excel in *some* field are far from new, but are socially helpful. Spearman pointed out at the British Association in 1928 that the theory of *g* left variations in memory capacity and numerous special abilities uncorrelated, and that, in consequence, "Everyone can be a genius at something." Perhaps the division in the Middle Ages between the study of the *trivium* and the *quadrivium* had also in part the function of giving self-respect to the less abstract performers in the way that Jensen suggests.

Such ameliorations of individual emotional conflicts, however, will not do as much good in the long run as a clarification of the general social philosophy on individual and group differences. We have to get used to the idea that statistically significant mean differences exist among groups and that these will usually be partly genetic and partly cultural. Lynn (1978), and Loehlin, Lindzey & Shuhler, (1975) report these for intelligence and the present reviewer for personality measures (1972). They have also been found for cities and for regions within a country, presumably due to migrational selection. Quite unnecessary heat and noise will continue to render would-be scientific debates on behavior genetics an interminable haggling until the value premises are frankly and clearly examined apart from the scientific findings. Politically, the American Constitution is founded on equality of rights and opportunities. No one supposes that such intelligent and scientifically informed authors in the Declaration of Independence as Jefferson and Franklin spoke of *biological* equality, or expected all to reach the same levels of intelligence. Today, however, in certain political groups, the term egalitarian no longer refers to the liberal philosophy of equal justice and democratic organization, but to a Watsonian belief that with suitable conditioning all can be brought to the same intellectual level. Elsewhere Cattell (1972), in examining the possibility of an objective basis for ethics, the present reviewer has pointed out, as have the sociobiologists with more detailed illustration since, that human progress, as in all evolution, depends first on variation, that is, on individual differences, genetic and acquired. An ethics of progress, rather than stagnation, therefore begins with an acceptance – and indeed a fostering – of individuation.

Jensen's second and third chapters, "Tests on Trial" and "The Drive for Equality," follow through on the above principles at the level of illustration found in practical educational and legal issues, though without the basic and intensive examination of principles per se mentioned above. These chapters suffice fully, however, to bring out the confusion rampant in the public and the law courts on the meaning of "discrimination," "justice," the Fourteenth Amendment, and the best organization of education to enable each individual to reach his full potential. Herrnstein's (1971) *IQ in the Meritocracy* attacks much the same issues as these chapters, and the convergence of these two independent-minded writers is encouraging.

For those journalists whose social inferences and psychological knowledge are equally erratic this book may cause as much tumult and denigration as Jensen's other books. For the graduate student and professional psychologist, on the other hand, it will take its place with his last two volumes as a solid and scholarly contribution. Indeed, in the area to which it professes to contribute there is today nothing to equal it, and the timelessness of its psychometric analyses is likely to make it a leading work of references for some years to come.

Editorial Note. *The following review, originally commissioned, and then not published, by The London Review of Books, appears in BBS in its entirety.*

by **Ann M. Clarke**

Department of Psychology, University of Hull, Hull HU6 7RX, England

Unbiased tests and biased people

Human beings differ – some of us might even say "vivent les différences." It would be a dull society indeed in which everyone was similar in temperament and talent; those of us who are not gifted athletically, musically, artistically may take pleasure in the activities of those who are. However, sometimes differences lead to problems, and all societies with compulsory education have in common the fact that there are very considerable differences in children's abilities to learn at school and among adults to profit from higher education. People vary in their ability to handle situations which demand certain kinds of complex and abstract reasoning, which leads to preferment in selection for entry into certain occupations. The distinction between workers by hand and by brain to which Marx alluded is enshrined in the official

census data in every country which attempts to obtain statistical information on its citizens. There may be profound divisions of opinion as to the causes of such differences, but that there are real, important social phenomena is undeniable. In Great Britain we have a problem in that perhaps some 15 percent of the population experience very considerable difficulty within our educational system. In the U.S.A. this situation is exacerbated by the fact that a disproportionate number come from an ethnic minority, representing only 12 percent of the total population, whose ancestors were slaves forcibly introduced from the African continent.

In 1904 Alfred Binet, a psychologist, and Théophile Simon, a psychiatrist, were commissioned by the Ministry of Education in France to devise a practical means for distinguishing between mentally retarded and normal school children so that the former could be quickly identified and provided with special education. They were responsible for producing the first intelligence test to be widely used in schools both in their own country and in others. In effect, they initiated the mental test movement, which has developed over the years in sophistication and influence, providing selection devices for various sections of the armed forces, industry, and education.

Mental tests, provided they were properly standardised, were shown to be a quick and relatively accurate way of predicting how effectively people would adapt to a variety of situations, including those demanding technical learning, within a recognised margin of error. Unfortunately and inevitably many of those taking these tests found themselves at the bottom of the hierarchy. The tests, in order to be useful, had to reflect real differences in adaptation to real situations. In societies like America, with a strong egalitarian tradition, it was inevitable, and proper, that attempts were made to reduce the differences. One of these attempts involved the accusation that the tests themselves were biased.

In Jensen's long, scholarly, clearly written book, the first chapter is devoted to the variety of criticisms levelled against psychological testing in all its aspects. These are extensively documented, and some of them make strange reading; for example, "The minority, disadvantaged student may incur educational damage by being subjected to current standardised testing methods." Organized opposition included a three-day national conference arranged in 1972 by the National Education Association on the theme: "Tests and Use of Tests – Violations of Human and Civil Rights." In the second chapter, the author moves on to a disturbing account of legal cases in which the plaintiffs' claims were that they or their children had been placed at a disadvantage as a result of being assessed on IQ tests which are culturally biased against Negroes. In one typical case the court accepted as *prima facie* evidence for test bias the plaintiffs' statistics indicating that Negroes comprise 9.1 percent of all schoolchildren in California but 27.5 percent of children in classes for the retarded, admission into which requires (among other things) a score below IQ 75 on an individually administered intelligence test. These undisputed statistics were claimed to support the charge of racially biased tests on the assumption that scholastic aptitude is equally distributed in all races, an assumption that went unchallenged. Various court cases resulted in legal decisions with far-reaching consequences for the practical application of tests. Most judgments have been based on the rarely questioned premise that the distribution of ability is the same in all subsections of the general population, which, of course, includes racial minorities.

In his preface, Professor Jensen states: "Many widely used standardized tests of mental ability consistently show sizable differences in the average scores obtained by various native-born racial and social subpopulations in the United States. Anyone who would claim that all such tests are therefore culturally biased will henceforth have this book to contend with."

The author argues persuasively that most tests measure something worth measuring, and that his exhaustive review of empirical research bearing on the problem of *bias* leads to the conclusion that the standardised tests currently used most widely, (and many less widely used ones, of which there are an astonishing number) are not biased against any of the native-born *English-speaking* minority groups on which the amount of research evidence is sufficient for an objective assessment. For most nonverbal standardised tests, this generalisation is not limited to English-speaking minorities.

The book is, among other things, a clearly written, well-ordered exposition of the theory and technology of psychometrics, which could be read with profit by those generally interested in this field of applied research, and recommended as an excellent text for students. It is, in addition, a mine of information on certain aspects of differential psychology, providing summaries of recent research on a host of topics, and includes updated empirical evidence on the reliabilities and validities of various instruments.

An unbiased person who has never heard of A. R. Jensen will find little in this book to suggest the author's position on the heritability issue. The author distinguishes between complexity and rarity of test items, and comes out in favour of the former, while acknowledging that performance on both types of test is highly correlated within a culturally homogeneous population (p. 640). He talks about various ways of viewing intelligence, its phenotypic character (i.e., the form in which it is overtly manifest) being the only one which can be measured, lists under causes of correlations a substantial paragraph on environmental correlations, and states that although there are methods in quantitative genetics by which we can analyse the correlation between two traits into two components – genetic correlation and nongenetic or environmental correlation – these methods have not been applied to human data (pp. 194–195). He devotes most of two pages (pp. 284–285) to environmental causes of instability in test scores, and (p. 569) accepts a social-cultural explanation for an important upward shift, particularly for preschool children, which is apparent when comparing the performance of children on the Stanford-Binet test in the 1930s with their performance in the 1970s.

So what is all the fuss about? In any case, do we need tests at all? These are two separate questions, which have become perhaps unfortunately related. The antipathy towards mental testing arises largely from the fact that most of the pioneers of intelligence testing, Alfred Binet apart, held strong hereditarian views and believed that an IQ score was a relatively direct measure of genetic potential, a philosophical position which persists in the minds of many advocates of testing to this day. Furthermore, in 1969 Jensen (1969) published in the *Harvard Educational Review* a long article in which he attributed differences in IQ and scholastic achievement in the main to genetic differences, and offered this as the major reason for the failure of the Headstart programme significantly to improve the learning abilities of young children on entry into school. A very large number of these children were black, and the implication was clear: their IQs and scholastic attainment could not be boosted. Alternative explanations were rapidly supplied by his opponents, but Jensen had himself arranged the scenario for the public outcry against mental testing which included the charge of test bias.

Perhaps if we abolished testing the differences would go away, or at least remain obscure. This seems to me to have been the hope of some psychologists who embraced Piaget's theory of development, which stressed the progress of all children through developmental stages determined by the essentially natural transactions of the child with his environment [cf. Brainerd: "The Stage Concept in Cognitive-Developmental Theory" *BBS* 1(2) 1978]. This is a process model of development which stresses what is common in children's progress, largely ignoring individual differences. Unfortunately, however, when various subgroups of the American population were assessed on Piagetian tasks, the disparities in performance across social class (and ethnic) groups was, if anything, greater than on conventional standardised IQ tests. Teachers are still going to have to cope with children whose capacity for learning varies enormously: the differences cannot remain obscure.

Since mental testing has become such an emotive issue, might it be wiser to leave assessment to teachers and employers? On page 173 Jensen presents some evidence indicating a surprisingly high relation between teachers' ratings of pupils for "brightness" and their tested IQs, the correlations being typically between .60 and .80. He goes on to discuss some well-known biases in teachers' ratings, which are unlikely to be present in properly conducted assessments by standardised tests, so perhaps the latter are a fairer way of assessing pupils.

So far as adult selection is concerned, the author offers a very clear account of three major philosophical conceptions of fairness together with the argument for and against each. These are: unqualified individualism (in which *all* relevant predictive information is used); qualified individualism (in which the same material is used but identification of the applicant's sex, race, and social status is withheld); and the quota system (which ensures as a matter of policy that members of disadvantaged subgroups are selected at the expense sometimes of better qualified but more advantaged applicants). Organizations must obviously be free to decide which of these philosophies best suits their needs within a particular political context. However, Jensen suggests that if standardised assessment procedures were to be abolished, industries and universities would almost certainly devise their own informal selection procedures, with no guarantee that these would be fairer to applicants.

Appropriately in a book about *test bias*, written by an author who sticks to his brief, some of the problems in interpreting and using psychometric data are not discussed, although the final chapter is about uses and abuses of tests. Nor are the causes of group differences evaluated, although they are in the background thinking of the protagonists. However, in conclusion, let us assume for a moment that the major reasons for subgroup differences within a society such as ours are large variations in the environmental backgrounds of native-born English-speaking citizens, although at present there is little acceptable evidence to substantiate such a position. Such an assumption would have very little bearing on the issues discussed in this book, and, in any case, no social philosopher has as yet been able to propose a viable method for overcoming the environmental deficiencies which many people (including recently Jensen) assume depress the potential functioning of a minority of a nation's most deprived children. It would be exhilarating to believe that in the near future parents will be demanding that their children be IQ-tested in order to qualify for special classes for the mentally retarded; this will only happen when and if the educational system becomes able to ensure that socially disadvantaged children are enabled to learn at the same rate as the more advantaged. The attempts to date are not encouraging. Let us not blame tests for biases in society. But those, including perhaps members of the legal profession, who believe that tests are biased rather than people, had better read this book.

by **Donald D. Dorfman**

Department of Psychology, University of Iowa, Iowa City, Iowa 52242

Test bias: What did Yale, Harvard, Rolls-Royce, and a black have in common in 1917?

A scholar who reviews issues and evidence on a question that has broad implications for public policy, and who often addresses his discussion to the nonspecialist, has a particular obligation to give an unbiased and scrupulous presentation.

Jensen's book contains a number of statistical misstatements – some that would seriously mislead the nonspecialist. For instance, Jensen asserts: "Path analysis is a method for inferring causal relationships from the intercorrelations among the variables when there is prior knowledge of a temporal sequence among the variables" (p. 336). The correct statement about path analysis is: "Path analysis is not a method for discovering causal links among variables from the values of correlation coefficients" (Fienberg 1977, p. 91). The problem is that Jensen wishes to infer causation from observational studies. But "mere observational studies can easily lead to stupidities" (Kempthorne 1978, p. 1).

Evaluation of Jensen's arguments often requires a precise understanding of the Pearson product-moment correlation – a fundamental statistic in many of his analyses and discussions. The author goes over that statistic in great detail for the nonspecialist, explaining that the Pearson correlation (ρ) "is a quantitative index of the degree of relationship between two variables" (p. 187), and that ρ^2 "expresses the proportion of *variance* [his italics] in Y that is predicted by or associated with X" (p. 190). Those statements are erroneous. The eminent probabilist Feller (1957) characterizes the correlation coefficient correctly: "the correlation coefficient is by no means a general measure of dependence between X and Y. However, p(X,Y) is connected with *linear* [Feller's emphasis] dependence of X and Y" (p. 222). This is no minor point. The correlation coefficient may vanish even if Y is a function of X. The fundamental problem is that the Pearson index of relationship is not invariant over monotonic nonlinear transformations of X and Y. Indeed, under some conditions such transformations radically alter the size of the correlation. Hence, unless the underlying regressions are known to be linear, theoretical or causal inferences based upon the Pearson correlation coefficient are of dubious value.

Jensen – a major publicist of Sir Cyril Burt's research (e.g., see Jensen 1972; 1973) – continues to cite Burt's suspicious data in his new book with no reference whatsoever to their questionable nature. Recent evidence has shown beyond doubt that Burt fabricated IQ data (e.g., Hearnshaw 1979), and such fabrications may have taken place as early as 1921 (Clarke & Clarke 1979). Hence, Burt's data are surely inadmissible as evidence in scholarly debate and discussion. Nevertheless, in support of his discussion of the distribution of intelligence presented in his chapter "The Distribution of Mental Ability," Jensen displays two of Burt's frequency distributions of IQ, one published in 1957 (see Figure 4.11, p. 80) and the other published in 1963 (see Figure 4N.1, p. 120). He also makes indirect use of Burt's questionable data in that chapter. There Jensen claims that "the study of identical twins provides good evidence that the environmental influences on IQ are normally distributed" (p. 119) and refers the reader back to his 1970 paper, "IQ's of Identical Twins Reared Apart" (Jensen 1970) for that evidence. In that paper, we find that Burt's discredited twin data constituted almost half the sample upon which Jensen's conclusion was based. I should also point out that Jensen's theory that general intelligence is approximately normally distributed and that achievement is markedly skewed is untestable. The skew of the corresponding empirical distributions is determined by the average difficulty of the items, and the kurtosis by the intercorrelation among the items, irrespective of any true underlying distribution. Jensen acknowledges this fact but still thinks that his theory is testable.

Jensen also cites (p. 359) Burt's research on delinquency and IQ from the old delinquent's *The Young Delinquent* (1925). In that work, Burt actually argued for subjective assessments based upon physiognomy: "To the observer who knows what signs to look for, the child's face, physique, and general deportment are always rich in significance. Physiognomy as a science has been much neglected" (Burt 1925, p. 413). One of Burt's classic physiognomic assessments can be found in that book: "In looks he was a typical slum-monkey. His sloping forehead, his diminutive snub nose, his prominent jaws and lips, were suggestive of the muzzle of a pale-faced chimpanzee" (Burt 1925, p. 302). Jensen also cites Burt's findings without attribution. At the end of his chapter 7, Jensen presents a model for intellectual development. In support of that model, he writes (p. 291): "The model is entirely consistent with the following summarization of research findings on mental development put forth by a group of British psychologists in a government report on secondary education." Jensen gives the primary source of that summarization as the "Spens report 1958" (p. 291). In fact, the well-known Spens report was published in 1939 (Board of Education 1939) and Jensen's quotation is found in part II of chapter III of that report. The report states (p. 120), "Part II of this chapter is based on a Memorandum prepared for the Committee by Professor Burt." That summarization should not, therefore, have been quoted by Jensen because of its untrustworthy source: Sir Cyril Burt.

To Arthur Jensen, the key question appears to be: "Are the criticisms of tests by blacks and their white sympathizers ill-founded and misdirected, or are they just?" (p. 51). He concludes that most standardized tests of mental ability are unbiased with respect to race and class for native-born, English-speaking people. He presents a plausible case that empirical validity and reliability do not vary greatly across race and class, but it is on the question of *construct validity* that his case fails. Jensen presented the principle that "if the items are to measure intelligence, they must possess certain abstract properties, described by Spearman as presenting the possibility for *eduction of relations and correlates*" (p. 127). I accept that principle. He also presented the principle that "The subject must first know the elements

of the test item and understand the requirements of the task for it to reflect the subject's power of eduction" (p. 130). I accept that principle also. It follows from those two important principles that the measurement of the bias of a test item on construct validity requires *objective* measurement of the frequency of exposure of the subgroup in question to the elements of the item and to the general task. Such a determination demands an objective assessment of the cultural environment, and such assessments of cultural environment are virtually never performed. Psychometric analyses of test items are, of course, irrelevant to the question of differential cultural exposure of various subgroups to the elements of the test items and to the general tasks. One must also assess motivation across subgroups: *motivation* is presumably not equivalent to intelligence. Accordingly, until objective measures of cultural experience are available and shown to be equal across subgroups, the bias of IQ tests for the assessment of intelligence remains indeterminate.

In 1917, an illustrious group of mental testers developed the Army Alpha – the first major group test of intelligence used to classify national and racial groups according to intelligence. On the basis of psychometric properties, those mental testers used items on geographical location: "Harvard University is in ____ "; Yale University is at ____ ." They used items on sport: "Lob is a term used in ____ "; "Slice is a term used in ____ ." And they used items on automobiles: "The Pierce Arrow car is made in ____ "; "The Rolls-Royce car is made in ____ ." I think that we would all – with the possible exception of Arthur Jensen – agree that those items were inappropriate for the measurement of the intelligence of blacks in 1917, irrespective of their psychometric or statistical properties.

The fundamental irrelevance of Jensen's statistical formulas to the issue of construct validity brings to mind the remark of Sir Peter Medawar, the eminent British biologist and Nobel laureate, that a "distinguishing mark of unnatural scientists is their faith in the efficacy of statistical formulas, particularly when processed by a computer – the use of which is in itself interpreted as a mark of scientific manhood" (1977, p. 13).

Acknowledgment

I want to thank Jacob O. Sines and Lorraine T. Dorfman for their constructive reading of these remarks.

***by* Douglas Lee Eckberg**

Department of Sociology, University of Tulsa, Tulsa, Okla. 74104

The problem of hierarchial thought in the work of Arthur Jensen

Whatever one's opinion of Arthur Jensen or of his beliefs about human ability, this massive book will be a source of material and arguments. Like a lightning-rod, it will draw fire, much like Jensen's more than a decade-old article that began the modern IQ controversy.

This is also a disturbing book, not for the information it presents, but for the inferences it makes, for what it implies – counterfactually – about the nature of human achievement. For the record, I find Jensen's treatment of technical aspects of bias very strong. But half the book does not deal with test bias at all. Rather, it is a painstaking attempt to legitimize the idea of a general intelligence. This is not in itself bad, and Jensen does make some strong arguments (for example, he makes a stronger case for the existence of a more or less general set of intellective processes than I thought he would be able to do). However, the point of this elongated treatment of the construct of intelligence is to portray humans as, in some simple sense, varying along a hierarchy of ability, from those at the low end to those at the high, on the basis of a single characteristic. It is this aspect of Jensen's work which has inflamed his critics, much to his apparent dismay and astonishment. In the short space allotted, I will sketch the reasons for this.

Jensen seems genuinely confused by his critics' harsh responses to his work and to testing in general, and when he speaks of them it is almost wholly in a manner which reads them out of the society of reasonable people. Thus we find, in what is little more than a subtle *ad hominem* argument, that "critics often try to ridicule tests" (p. 4); "small but vociferous groups . . . have waged propaganda campaigns" (p. 16) against testing; there is an "antitest syndrome" (p. 18) in the air; critics have claimed that intelligence is "just a fiction invented by the 'establishment' to justify inequalities and perpetuate various forms of social injustices" (p. 175); and that some theorists "would like to have us believe" (p. 346) that occupational level is not causally related to IQ at all. Jensen insulates himself from such people by adopting the cloak of positive science. If the book is overly long, it seems to be because he believes that he needs merely to hit us over the head with enough formulae, definitions, citations, facts, and inferences, and we will be forced (if reasonable) to agree with his overall position. If his critics do not agree with it, especially if they make statements, on occasion, that seem silly when contrasted to the relevant literature (especially as he often presents straw-man arguments to attack), then ipso facto they cannot be reasonable.

But are the sources of the acrimonious response to Jensen's position so obscure? They are not. Rather, they can be found in the common meaning of intelligence, as it exists in both social thought and in much of the psychometric literature. By tradition, "intelligence" has been presented as that characteristic which separates humankind from the beasts and allies them with God and the angels. This is most clearly found in the thought of the Elizabethans (e.g., Tillyard 1944, pp. 71–74), where "reason" (consisting of "understanding" and "will") was the prime human characteristic. To deny one's ability to reason was to deny a large measure of one's humanity. This idea has remained almost unchanged. For example, with Herbert Spencer's (1899, vol. 1, p. 4) evolutionism, the highest mental attributes were "reason," "the feelings," and "will," existing in humans alone. Francis Galton's (1892) view of intellect, while marked by an inconsistency between an ability and an abilities approach to eminence, generally opted for the former, along with "zeal" (read: "will"). In the U.S., Lewis Terman demonstrated a clear reverence for the idea of a general intelligence, which would be *the* variable of note in the assessment of human potential. This is evidenced by his adulation of gifted children, and by both his truly disparaging remarks concerning low-IQ children and his successful attempts to arouse political pressures in favor of laws calling for the sterilization of low-IQ people (see Terman 1916, pp. 91–92; 1917). As I have shown elsewhere (Eckberg 1979, chs. 6–7), Terman's colleagues for the most part shared his concerns. Politically, the concept of intelligence has been used to justify slavery (Jordan 1968, pp. 304–311, 440–457), to halt immigration, and to deny education.

Where does this leave us now? Certainly not in the same place as a half century ago, though there are similarities. The most basic similarity has to do with the postulation of a central mental faculty which separates those with ability from those without. Jensen and his cotheorists acknowledge processes independent of *g*, and acknowledge that a number of occupational tasks can be accomplished with a minimal amount of it. Yet, *g* is continually put forward as the *sine qua non* of achievement. In the present work, Jensen in several places admits to the rather modest relationships between IQ test scores and common indicators of achievement. Yet, throughout the bulk of the book, he treates the scores *as though* they were of supreme importance. What is the evidence? It is too massive to spell out here, though various writers (e.g., Eckberg 1979, ch. 4; Bowles and Gintis 1976; Jencks 1972; McClelland 1974; Wallach 1976) have gone into some detail on it. Basically, IQ scores seem not to be negatively correlated with performance on any tasks. They are modestly correlated with various indicators of performance on a large number of tasks, and seem more strongly associated with achievement on some more "intellectual" tasks (though Jensen does not mention the counter-evidence on this point). When factor analyzed, IQ tests – which are largely made up of short puzzle, multiple-choice items included on the basis of intercorrelation – commonly show about one-third to one-half of their variance loading onto a single factor. "Purer" tests correlate less well with achievement than do less pure tests.

When such a lengthy exegesis is so painstakingly constructed with the express purpose of demonstrating the centrality of a construct

which accounts for such a relatively small amount of the full range of human achievement, with people portrayed (with some hedging) as ranging along a simple hierarchy of ability, then there certainly is room for dispute. Reasonable people need not accept Jensen's conclusions. Instead, his basic theme can be disputed. Considering those conclusions, it is not unreasonable even to question his underlying assumptions regarding the nature of social stratification. Therefore, *Bias in Mental Testing,* rather than sounding the death knell for debate, should serve as the springboard for a new round of controversy.

by **Bruce K. Eckland**

Department of Sociology, University of North Carolina, Chapel Hill, N.C. 27514

Competent teachers and competent students

Bias in Mental Testing strongly supports the administrative use of mental tests in postsecondary education and employment (but less so in elementary schools). The book poses a direct challenge to those of us who assume that verbal tests are much more biased (comparing blacks and whites) than are nonverbal tests, that the effects of "Pygmalion in the classroom" [see Rosenthal & Rubin: "Interpersonal Expectancy Effects" *BBS* 1(3) 1978] are the gospel truth, and that the advantages of ability grouping in the elementary school outweigh any disadvantages.

As I am an educational sociologist, my attention focuses mainly on school and college, which are also the main subjects of the book. There are six topics on which I have comments. I will begin with curriculum tracking and minimum competency testing in the high school, then move on to the college years, and then back again to the schools, where I will deal with the relationship between competent teachers and competent students.

Tracking in high school. There is one particular event in the educational process which Jensen says little about but which is probably more powerful than anything else in the cycle in determining who gets educated; it is also closely related to mental testing: curriculum tracking in high schools. Jensen argues that the issues of tracking could be made irrelevant at the high school level if students were given more guidance and freedom in selecting the courses they wanted. In a sense, I think he is right, but he does not deal with the formal manner in which the high school curriculum is structured, or with its causes and consequences.

On the basis of the National Longitudinal Study of the High School Class of 1972 (the NLS is a sample of about 22,000 seniors in over 1,000 schools, sponsored by the National Center for Education Statistics) 78 percent of those in the high school academic track entered college within four years after graduation compared to only 17 percent of those in the vocational track and 33 percent of those in a general program. Track placement was found to have stronger direct effects on college entry than did tests scores, high school grades, or social class background (Thomas, Alexander, & Eckland 1979).

Some writers have claimed that who *enters* the academic track is largely a function of family background. In the NLS, however, we found that mental test scores were substantially more predictive of track placement than were either social class or race. The "direct effect" of each of these variables on being placed in the college preparatory track was .16 for social class (controlling for race and ability), a *positive* .12 for being black (controlling for class and ability), and .52 for ability (controlling for class and race). Mental test scores far outweigh other factors in the placement process.

There is probably no single event in the educational cycle that has a greater bearing on more people, and we continually seem to ignore it (Jensen is not the only one).

Minimum competency testing. I must take issue with Jensen on his discussion of the competency testing movement in high schools. Two of the reasons he offers against the practice of giving "certificates" instead of diplomas to graduates who do not pass the test are: (a) using an imaginary line between "minimal competence" and "incompetence" stigmatizes the individual; and (b) the schools could use better means to teach students what they need to know "even if, for some pupils, it means repeating a whole grade" (p. 725).

With the idea that the labels would be stigmatizing, elsewhere in the book Jensen makes the point that there has not been much research support at all for the numerous proponents of labelling theory in the schools (his main reference is to studies on teacher expectancy effects). On this I agree with the research literature. It is unlikely that it would make much difference *on these grounds* in the long run for students who will be awarded "certificates" instead of "diplomas." The vast majority of those who fail the competency tests certainly know how poorly they can read, write, or calculate the cost of groceries. That these labels would seriously affect them requires documentation that is not likely to be forthcoming.

Jensen's reference to repeating grades in school might be defended if schools would really do it. However, it is very unlikely, given the current state of affairs. Moreover, if social promotion went out of style, students who were held behind in grade level would tend to drop out of school before ever receiving either a diploma or certificate, in which case the argument becomes moot.

My main claim against Jensen's position, however, has nothing to do with the above but with something he ignores. That is, in my estimation, the primary value of minimum competency testing has been in the introduction of many new remedial education programs in the schools. Students do not take their first tests as seniors but usually one or two years earlier. The proportion of those who fail (at least in Florida and North Carolina) has been greatly reduced by means of such programs. For the first time in public education, our high schools are making themselves accountable (at least to some extent) for the competence of their students and are attempting to do something about it. How can Jensen argue that this "is surely one of the most futile proposals to come along in public education in many a decade?"

Race, test scores, and college attendance. Jensen is well aware of the race issue in mental testing for college admission, and of the fact that there is much pressure on colleges today to ignore the test scores of blacks (in the name of either affirmative action or reverse discrimination). Jensen also recognizes that one of the main arguments of the liberal egalitarians in support of this practice is that college admissions tests are not very predictive of the grade performance of blacks after they are admitted, and Jensen presents much evidence to the contrary (as did Breland in a recent ETS report, 1979). However, he says almost nothing about race differences in the effects of test scores on who goes to college or who graduates. Both matters are of concern to egalitarians.

Although low test scores cannot stop anyone from going to college (I do not know of any state that does not have an open admissions policy at some institutions), most high school graduates probably have a fairly good idea of how capable they really are and whether or not they should go to college. Once a student is admitted, test scores probably are not only predictive of college grades but of whether a student graduates or becomes a dropout. I will not review all the evidence on these issues, but will report a few recent findings from the NLS (based mostly on unpublished data).

In predicting who goes to college, test scores are surprisingly accurate for both blacks and whites. By October 1976, 53 percent of the high school class of 1972 in the U.S. had enrolled in an academic program in college somewhere. The attendance rates of those in the low ability quartile (blacks versus whites) were 36 and 21 percent in the middle two quartiles 70 and 51 percent, and in the upper quartile 93 and 82 percent. Not only were blacks more likely than whites to have gone to college (within each ability quartile) but test scores were about equally predictive of who goes to college for both populations.

For those who entered an academic college program in 1972, test scores were also a good predictor of who had dropped out by 1976 without graduating. The dropout rates markedly differ by ability quartile, with whites again being the more "disadvantaged" within each quartile. The rates (no college degree and no longer enrolled) for those in the lowest ability quartile (blacks versus whites) were 56 and 68 percent, in the middle two quartiles 39 and 45 percent, and in the upper quartile 18 and 25 percent. The NLS data would seem to confirm Jensen's conclusion that mental tests are not only valid predictors of

college achievements but are not biased against blacks (see Eckland & Alexander 1980).

On the IQ of teachers. Jensen writes extensively on the use of ability test scores in predicting the performance of adults in different occupations. One occupation to which he unfortunately gives scant attention is *teaching*. The old cliché that "if you can't do, you can teach" is not dead, as anyone who is familiar with the curriculum choices of most undergraduates can confirm. Moreover, with the movement of an increasing number of college women and blacks [out of teaching] into more lucrative professions, the intellectual caliber of teachers is probably getting even worse.

Jensen does not deal with this issue but does note that there have been attempts in the courts to bar the use of test scores, like the National Teachers Examination, in the selection of teachers in some states and communities. The plaintiffs, not surprisingly, are usually blacks who tend to score much lower than whites on the tests. An issue which Jensen does not discuss is just what difference good and bad teachers really make in the schools. Even though Jensen makes good use of data from the largest school testing study ever conducted in the U.S., the Coleman Report (Coleman et al. 1966), he overlooks what this study had to say about the effects of schooling on the cognitive performance of students. What most researchers in education are aware of, but seem to want to forget, is one of the central findings of the study. That is, of all the various measures of school resources and teachers that were examined, the one thing over which schools have some control and that explained the most variance in pupil performance was a thirty-item vocabulary test *administered to the teachers*.

Of course, one could argue that the correlation between pupils and teachers on verbal tests is not causal but spurious, that is, due to the tendency that higher ability students (even blacks) tend to come from somewhat higher socioeconomic backgrounds and that their parents are more likely than others to see that they are enrolled in a good school with competent teachers. However, when controlling for the social background of students, the report concluded that teachers still had a marked impact on student performance, especially that of blacks at all grade levels.

To me, at least, this suggests two things. One deals with Jensen's proposal that IQ tests administered for purposes of educational research should be restricted to a "relatively small percentage of the total school population" and "with small samples from each grade in school" (p. 721). I would argue that we also need student samples in the classes of different teachers at each grade level. Numerous researchers have been arguing that schooling exerts much more impact on the cognitive development of students than studies like the Coleman Report have shown, but one of the problems is that most of the variance in the quality of teachers or the use of school resources generally lies *within* rather than between schools. Thus, in order to find out what really is happening in the schools, studies on the cognitive development of children must move down to the level of the classroom, and *that* is where we need access to much better data.

My other point (and this needs no response) is that I can think of no place in the entire world of work that is more depressing (in terms of outcomes) and in need of change than the teaching profession. My reasons are based not only on the low status of the profession but also on the low intellectual caliber of so many of those who enter and *stay* in it, and the situation could be getting worse. Also, the consequences for contemporary society could be becoming more serious because of the increasing reliance of parents on the schools in the rearing of children.

When Florida instituted minimum competency testing in the high schools, its intent was to hold teachers accountable for the education of their students. However, in Florida, as elsewhere, it is the students who eventually will be penalized, not the teachers. There are ways of firing a teacher who violates the most serious rules of conduct, for example, one who rapes a student. But there is no way of removing an incompetent teacher. Given the manner in which our schools operate, there is not even a way of keeping an incompetent teacher's salary from being raised or of increasing the salary of a competent teacher. Merit is going out of style in many areas of employment these days. Unfortunately, it never was much in style in the teaching profession.

by **Judith Economos**

Renaissance Studio, Scarsdale, N.Y. 10583

Bias cuts deeper than scores

Professor Jensen is famous for being a respected scholar who defended the proposition that the population of white Americans is significantly better endowed by nature than is the population of black Americans (Jensen 1969, 1973). He did so on the basis of divers data, the most persuasive of which were almost certainly tainted (Kamin 1974, Jensen 1974) and a good deal of statistics, much of which was evidently inapplicable.[1] Since neither population has much genetic integrity, the proposition might in any case seem prima facie implausible.

In *Bias in Mental Testing*, Jensen is not concerned with promoting a genetic explanation for the persistently obtained one-sigma difference in IQ score distributions of black and white populations (although it is detectable that he still favors that hypothesis). He is concerned with showing that whatever the explanation may be, it is not that there is a built-in antiblack bias in the tests. We ought to consider *Bias* independently of the genetic explanation; and as there are no commonly accepted facts bearing on that issue, we leave to others a battle fought with eloquence, epithet, and indignation.

Some, contemplating the one-sigma difference, will find it difficult to believe that it can be explained by the existence of an inherent difference in intellectual ability between two such large and heterogenous subpopulations of Americans. These people will look for other explanations of the difference. For example:

It can be denied that the tests measure a stable entity to be called 'intelligence.' If so, the tests lose all significance and have only a temporary, diagnostic utility.

It can be denied that the tests all measure the same, general, unitary capacity. In this case, test results are not generalisable, or at best measure only a narrow, specialized academic "knack," and not at all the prized range of mental powers we mean by 'intelligence.'

It can be denied that the tests measure any objective trait at all: the scores are just artifacts of the tests, and the same subjects could be ordered differently by changing the tests. Alternatively, it can be argued that the tests simply measure the degree of immersion of the subjects in white, upper-middle-class culture.

It can be asserted that the tests do measure something objectively, to wit, not innate intelligence alone but also the pervasive deprivation of the black population in whatever those environmental features are which encourage the development of abstracting intellect.

Bias considers all these points, in various forms, and defends the tests – more or less, depending on the test – against them. Jensen offers statistical arguments that mental tests measure, among other things, a common, stable, unitary, objective trait reasonably called 'general intelligence,' and that the admitted and considerable effects of socio-economic deprivation can be "partialled out" of the results, leaving an unexplained racial difference.

In order to present his arguments, Jensen offers an excellent first course in psychological statistics and an interesting glimpse of the techniques and standards of mental-test construction. To have done these things lends credit and credence to his efforts. After considering his patient arguments, bolstered by so much, well, biastatistics, I am persuaded that the better tests are not systematically biased against blackness, and that they do manage to address a common general ability plausibly thought of as intelligence, which is (with some vivid exceptions) relatively stable in individuals. I cannot see why anyone should object to this very loose statement; in any case, I think it is true. It does not claim that anybody's intelligence is accurately determined as fixed for life, or beyond improvement. It corresponds at its harshest to the proposition that people who score poorly on these tests will almost always find it harder, for example, to follow advanced mathematical reasoning, or quickly to extract the meaning from a scholarly paragraph, than will people who score well on them. (Nor does it say that these are the most desirable or advantageous human abilities.)

Where Jensen may be walking on quicksand is in relying on various "socio-economic scales" to subract blackness from test results. While being black is not an ineffable and mystical property, it is still fairly

obvious that being black and poor in America is not yet the same thing as being white and poor. It cannot be impossible to measure this difference, but I do not believe it has been done. The mental tests probably are not biased against blackness. It scarcely matters, for society still, observably, is; and that bias compounds itself.

If we assume that the test organizations desire to make their tests as widely used as possible, and note that highly motivated people keep them well informed about possible bias in their tests, and accept that there are techniques for identifying kinds and tendencies of bias, it is reasonable to suppose that the organizations are making their tests less and less biased. But whether or not one accepts these reasonings, it is a misdirection of energy to attack the tests, especially given what *Bias* tells us. There is an abundance of evidence, independent of and prior to test results, that in this society blackness cripples – on the whole, and with striking exceptions. Moreover, Christoper Jencks (1979) has offered evidence that test scores are really quite unimportant in determining "who gets ahead" in America; again, therefore, our energies would be better spent on equalizing people's opportunities than on searching for ways to jigger the tests to equalize scores. (*Bias,* as it happens, gives a detailed discussion on techniques and consequences of compensating for certain kinds of known test bias – as well as of "compensating" for lack of desired bias.)

Bias in Mental Testing has accomplished several praiseworthy tasks. 1. It has made it very clear that certain intuitive judgments about bias in test items are wholly unreliable, and often foolish. 2. It has offered good reason to believe that the better mental tests are not themselves systematic artificers of their unwelcome results. 3. On the contrary, it has argued that the tests are, within their limits, almost certainly less capricious and unfair than the uncontrolled judgments of teachers, interviewers, or even parents. 4. It has therefore recalled our attention to the need to look elsewhere for causes of the difference in test score distributions.

Intellect is like talent or beauty in a number of ways. It seems to run in families – but unreliably, and with spectacular exceptions. It produces assortative mating. We have no idea what felicitous confluences of variables cause it. It can be (unilluminatingly) measured; however, no-one agrees on its definition. And because we prize it greatly and accord it privilege, though it is a completely undeserved gift, it is at all times and among any company a touchy subject. Intellect is, however, further like talent and beauty in being too bright and sweet a thing to draggle in the muck of racial competition and struggle for advantage; nobody (we all agree) should ever be held back from developing such excellences as may be his. The tests (Jensen says, and I agree) are not holding back black children.

But something is.

Note

1. See, e.g., Lewontin 1970. A collection of objections on this point, including the Lewontin, and a reply to Lewontin by Jensen, is gathered in Block & Dworkin 1976.

by **Robert A. Gordon**

Department of Social Relations, Johns Hopkins University, Baltimore, Md. 21218

Implications of valid (and stubborn) IQ differences: An unstatesmanlike view

More than a decade ago, Jensen (1969, p. 82) reviewed then current environmental hypotheses that might explain the average IQ difference between blacks and whites and found them so weak that he advanced the "not unreasonable hypothesis that genetic factors are strongly implicated" (see also Jensen 1973). Despite numerous denunciations of Jensen's thought we are no closer now to an environmental explanation of this difference, which still stands at its World War I value – about 18 IQ points on the Stanford-Binet metric, where the white mean is 101.8 (Brigham 1923, p. 80, weighting officers at 12 percent in accordance with a modern infantry division's 11.2 percent; Johnson 1948; Gordon 1976, Table 7; DHEW 1976, Table 2; Department of the Army 1976; Terman & Merrill 1960, fig. 4).

Tuddenham (1948) has argued that white IQ test performance improved substantially between World Wars I and II. In view of the constant racial difference, this would imply a corresponding gain for blacks, and hence some grounds for optimism. Before Tuddenham's data are read as evidence of important gains in IQ, however, we must consider that photographs show World War I testees sitting on the floor, sometimes wearing overcoats; testing was often rushed or conducted under adverse conditions; and familiarity with ability tests was then at its historic minimum (Yerkes 1921). World War I soldiers were probably undereducated for their ability, too, whereas this would have been less true in 1943 when white enlisted men averaged two years more schooling (Tuddenham 1948). Performance gains, for example, in literacy are easier to obtain under such favorable starting conditions. More than likely, better performance by World War II soldiers on World War I tests in Tuddenham's study reflected paper gains due to "test sophistication" (Jensen 1980, pp. 589–591), better testing conditions, and perhaps education, rather than real gains in either fluid or even crystallized intelligence (on these terms, see Jensen 1980).

It also seems improbable that environmental improvements for blacks have remained perfectly synchronized with relative gains for whites throughout the history of ability testing. However, the racial mean IQ difference has stood equivalent to 1.1 white standard deviations (16.4 points) during World War I; during the twenty years following World War II (Shuey 1966, p. 503); near the start of the civil rights decade (Coleman et at. 1966; Jensen 1980, Table 10.3, grades 6 and 9); and at the end of that decade (DHEW 1976, Table 2). Unless there have occurred perfectly correlated secular trends in real IQ for both races, these data suggest remarkable stability for group means over a sixty-year period that saw mass migration from South to North, rural-urban shifts of large magnitude, changes in occupational and especially educational attainment, the Great Depression, increases in real income (until recently), television, wartime drafts, GI Bills, racial spurts and lags in relative socioeconomic standing, a gain in median family income for blacks from 37 percent to 62 percent of the white figure between 1939 and 1974 (Okun 1976), school desegregation, and billions spent for compensatory education (McDill, McDill, & Sprehe 1969, pp. 26–33).

Moreover, the required hypothetical gain in IQ for the black population between World Wars – a period of relative neglect, benign or otherwise – would stand in puzzling contrast to the lack of change produced by more intensive and deliberate efforts under policies begun in the mid-1960s (e.g., McDill et al. 1969). Data from Coleman et al. (1966) show no decrease in the racial gap in the course of schooling (Jensen 1980, Table 10.3), just as data for whites show no average change over the school career for children from different social classes (McNemar 1942, Table 10).

Although Plomin and DeFries (1980) report that within-race heritabilities from recent studies are "closer to .50 than .70" they also state that they "know of no specific environmental influences nor combinations of them that account for as much as 10 percent of the variance in IQ" (pp. 21–22). This would include Burks's (1928) well-known multiple correlation of .35 if allowance were made for shrinkage and its stepwise derivation.

Indeed, in the case of Larry P. (see Peckham 1979), where plaintiffs attributed overrepresentation of blacks in special classes for the retarded to use of "biased" IQ tests, the defense was hard-pressed to adduce plausible environmental explanations of the race difference in IQ means. The plaintiffs' strategy had been to urge the following false dilemma upon the judge: either IQ tests were biased or there must be genetic differences in IQ between blacks and whites. Judge Peckham (1979) found environmental arguments of the defense unconvincing – thereby concurring with Jensen – and, opting in favor of the less unattractive alternative of the false dilemma, pronounced the tests "biased" (Gordon, in press a). Each diagnosing by exclusion, Peckham and Jensen diverge only in the final elimination of alternatives, and it is clear that "test bias" has become the last specific refuge for determined environmentalism. Although no direct evidence for the genetic hypothesis exists, the completely environmental alternative becomes increasingly hypothetical as we are faced with accounting for "quite large race differences in IQ by very weak casual factors, as

judged by . . . effects . . . on IQ *within* races" (Jensen 1978, p. 22).

If an evironmental remedy were conveniently at hand, Jensen's book would be hailed as a demonstration of the need for applying it, that is, for taking the IQ difference seriously. But as Cronbach (1975, p. 4) observes, "Jensen was right about the failure of compensatory efforts, inasmuch as even now we have no compensatory method, reproducible on a large scale, of demonstrated value." Consequently, although Jensen's book barely touches on genetics, it will be viewed in the context of its relation to the environmental explanatory vacuum, and reactions to it must be read not only in terms of its definitive verification of the construct validity of stubborn phenotypic IQ differences – bad news enough – but also in terms of its inevitable implications concerning the likelihood of a purely environmental remedy.

Stubborn phenotypic differences, whatever their cause, promote exactly the same everyday outcomes as genetically based differences. Nevertheless, intellectuals regard genetic explanations with more dread than seems appropriate. Odd for a democracy, the public concerned is neither informed as to its options nor consulted as to its opinions. Genocide is often cited as a risk, but in practically all instances in this century the victims were apparently *higher* in average mental ability than either their oppressors or those permitted to survive (Gordon, in press b). Some observers appear to confuse gradual change in frequency of certain genotypes within a population with genocide. Large IQ differences and changing demographics place racial-ethnic groups on a sociopolitical collision course; before leadtime and goodwill are squandered altogether, it might behoove us to begin giving the various potential scenarios reasoned consideration. If tenured academics will not undertake this currently thankless task, who will?

by **Donald Ross Green**

CTB/McGraw-Hill, Monterey, Calif. 93940

Achievement test bias

Jensen's *Bias in Mental Testing* is a large and useful book. In an effort to prove that mental tests are, by and large, not biased he has presented a thorough and competent review of a very large body of material. Nevertheless, in spite of its length, the book is not fully comprehensive and inclusive in its treatment of the domain he sets out to evalutate, namely "standardized tests of mental ability – IQ, scholastic aptitude, and achievement tests" (p. ix).

The last of these, achievement tests, get much less attention than the rest. This may be due in part to the fact that a great deal of the work in this area has appeared relatively recently, since about 1975. In fact, it is probable that much of the book was written originally before then and subsequently brought more up to date in an unsystematic way as matters happened to come to Jensen's attention. For example, he discusses the California bill banning group IQ tests in schools as being before the legislature (p. 32), even though that law was passed in August 1975. In the next paragraph, he quotes Judge Peckham's decision in the Larry P. case. This decision did not appear until October 1979.

Another reason seems to be that Jensen was primarily concerned with IQ or aptitude measures, not achievement, as he wrote the book. Thus, a criticism of the book is that so strong was Jensen's concern with aptitude measurement that he gives achievement test bias inadequate attention. In fact, he does not really consider it a separate problem and draws no conclusions about achievement test bias separately from other mental tests. That is not to say that Jensen goes along with the proposition offered by some that aptitude and achievement tests are one and the same; rather, he devotes some space to the distinction between the two. He starts by noting that aptitude tests always measure performance that can be labelled achievement. He then goes on to point out a series of characteristics or features of aptitude tests which achievement tests do not share because he wants to show that aptitude tests are not merely achievement tests. The point of view, just as it is throughout the book, is that the measurement of aptitude is what is important.

However, one can usefully reverse Jensen's approach and point out that, while often differences in achievement test performance do partly reflect differences in aptitude, achievement tests are not just a kind of aptitude test. Although Jensen does not do so, one can infer from either approach that there are differences in the study of bias in the two kinds of tests. Since, generally speaking, criterion measures do not exist for achievement tests, for the most part only internal characteristics of achievement tests have been examined for validity and for bias. One consequence is that attempts to reduce bias in achievement tests have commonly used procedures which in effect assume overall validity of the test for all groups (Angoff 1975; Green 1975).

Most of the work on reducing bias in acheivement tests is not discussed in this book (reviews of this material can be found in Merz 1978 and Rudner 1977). It is noteworthy that most workers in the area have started with the assumption that the tests are biased to some degree. Some of these reports indicate that the identification and elimination of biased items can reduce achievement test bias at least a little (e.g., Ozenne, Van Gelder, & Cohen 1974; Green 1976). In other words, even though the amount may be relatively small, there is enough bias for it to be measurably reduced. Omission of this material reduces the comprehensiveness of Jensen's work. Of course, it is a bit unfair to criticize the author of a densely packed 800-page book for not including what could be considered an extra topic. However, I believe that a consideration of this material would require a modification of his conclusion that achievement tests are not biased.

But Jensen seems quite sure that test bias is negligible, and so I am not sure he would find this additional evidence convincing. There is obviously room for differences of opinion about what is negligible, and one's starting point will influence those opinions. Admittedly, I start out with the conviction that mental tests are biased to some degree (presumably small), whereas in this book Jensen sets out to prove they are not. His case is strong in many instances but it does appear to me somewhat less conclusive than he asserts. For example, he reports on thirty studies of the effects of the race of the examiner (pp. 596–603). Most (ten) of the sixteen "adequate" studies showed none, but three did produce effects which he then describes as "practically negligible or inconsistent." The fact is, that some studies found effects, and therefore there are inconsistencies. It seems to me preferable to try to find explanations for these inconsistencies rather than to draw a conclusion based on the majority of the studies.

Jensen also rejects some other propositions that seem to me plausible – albeit not well demonstrated – which point to the possibility of bias. One is the theory that some bias may pervade tests in a way that affects items almost equally (p. 574), so that it is not detected by interactions within the test. Another proposition, which could account for such pervasive bias, is that the entire testing situation is biased against some groups, leading to faulty conclusions about their competence (see, for example, Epps 1978; Hall 1978; Hall et al. 1977). This amounts to asserting that the concept of standardization is faulty, and while one may not find this an easy proposition to deal with, it cannot be merely rejected out of hand.

These are examples of issues that are still sufficiently open to preclude a final conclusion that mental tests are not biased, even though Jensen has indeed made it quite hard for anyone who respects evidence to claim that tests have proved to be substantially biased. As always, more work needs to be done.

Those who would do this work could do worse than beginning with *Bias in Mental Testing,* since the most impressive aspect of the book is the wide range of methodologies for the study of bias which it describes in detail. As Jensen himself suggests, he has written a textbook on methodology within the book. Nevertheless, here too there is an important omission, namely, approaches which use latent trait or item response theory. It seems almost certain that most of the work on bias that will be done in the immediate future will rely on latent trait methodology. Nevertheless, anyone interested in the topic of test bias will find the book an essential and fruitful source of information. Notwithstanding Jensen's clear and pervasive viewpoint, it should be possible for anyone to make good use of the material even if they do not agree with all his interpretations of the data.

by **Gordon M. Harrington**

Department of Psychology, University of Northern Iowa, Cedar Falls, Iowa 50613

Criteria of test bias: do the statistical models fit reality?

I suspect that the majority of critics will find fault with this book on substantive or strategic grounds while yielding to the weight of the arguments with respect to technically defined statistical or psychometric bias (e.g., Gould 1980). Even that ground, however, is not so easy to defend.

Mental testing is undergirded by a sophisticated quantitative methodology. For those without the necessary technical psychometric expertise, that methodology can be so cryptic that sometimes it is easy to believe it is irrelevant to reality. For those with the necessary technical psychometric expertise, that methodology can be so absorbing that sometimes it is easy to believe it is reality. Many criticisms of tests fall in the former category. *Bias in Mental Testing* tends toward the latter. Having correctly defined test bias as systematic errors in either predictive validity or construct validity, Jensen then offers detailed definitions or criteria of test bias usually cast in statistical terms. In some cases these assume the adequacy of the underlying statistical model without addressing the substantive question of the relation between the model and reality. An example from each of the major categories – predictive validity and construct validity – will illustrate.

Jensen introduces a new definition of predictive test bias – group differences in the parameters of the regression of the criterion on *estimated true scores*. The slope of this hypothetical equation is obtained by applying the correction for attenuation to the observed regression coefficient, thereby presumably adjusting for unreliability in the test. This is a correction in the predictor. Predictor correction has never been considered acceptable psychometric practice (Guilford 1954; Horst 1968; Nunally 1967). In practice, prediction can only be based on fallible tests and on observed data. The "correction" (statistical adjustment) estimates the relationship that would exist if the test were replaced by a hypothetical test with the same true score component but with zero error variance. As a problem in the philosophy of science it is not at all clear in this context what we even mean by a perfectly reliable test or by true score (see Rozeboom 1966). The actual question of test bias is one of the existence of systematic error in real tests. To define bias in terms of a hypothetical test devoid of variable error assumes that the statistical model transcends the empirical reality. Moreover, if a test is indeed biased, we don't know how to estimate the true score.

My comments assume a classical test model (e.g., Gulliksen 1950) in which true score is a construct with meaning which is independent of the test. Some psychometricians prefer an axiomatic approach (Lord & Novick 1968), in which true score is stochastically defined as the long-run average of the test scores. In this framework the statistical model fits by definition rather than by assumption. With stochastic definition a systematic error does not result in a biased test provided that the criterion measurement is subject to the same systematic error. Then, group differences affecting both test and criterion measurement are defined to be part of the true score. Debate on bias issues assumes substantive meaning. Under classical theory it is possible for both test and criterion to be equally biased, since true score has independent meaning.

For predictive validity, bias may be reduced to a matter of definition with a stochastic model, in which case the book is an exercise in triviality. With a classical model, under Jensen's definition for predictive validity bias, the criterion for bias is its presence in a test which does not even exist!

Turning to construct validation, group by item interactions are quite properly taken as the sine qua non of bias. Although a number of methods of assessing such effects are presented, all, of course, are conceptually rooted in the analysis of variance (ANOVA). How biased must a test be before it is a matter of practical significance as well as statistical significance? It is and has been argued by Jensen and by those he cites that, even though statistically significant, group by item interactions are usually small relative to racial effects and therefore unimportant. He introduces his own index for such a comparison, the Group Differences/Interaction Ratio:

$$GD/I = \frac{\text{Group } MS/\text{Subjects } MS}{\text{Group} \times \text{Items } MS/\text{Subjects} \times \text{Items } MS} = F_G/F_{GI}$$

or other expressions thereof. Since a test is unbiased unless there are group by item interactions, this index provides an alternative statistical representation of the relative magnitudes of effects in the appropriate analysis of variance. No objective rationale is offered or can be offered for interpretation of this index

What does the GD/I index represent? Before I provide a general answer, some specific statistical comments are in order. Consider the expected mean square in an ANOVA examining Groups (race), Items, Groups × Items and Subjects:

$$E(MS_R) = \sigma_e^2 + \sigma_{S(G)}^2 + N_S\sigma_{GI}^2 + N_I\sigma_{S(G)}^2 + N_IN_S\Sigma\gamma^2/(N_G - 1).$$

(I have expressed the last term in more expanded form.) The relative contribution of different components to the expected mean square is obviously dependent upon the sample size Ns so that GD/I and hence the judgement of bias varies with sample size. Now consider the expectation of the last component of variance, $E(\Sigma\gamma^2/(N_G - 1))$. Where γ is a random variable sampled from an infinite population of possible values, $E(\Sigma\gamma^2(N_G - 1)) = \sigma_G^2$. But race is a fixed effect so that $E(\Sigma\gamma^2/(N_G - 1)) = (N/N_G - 1)\,\sigma_G^2 = 2\,\sigma_G^2$ for a target population of two races. Thus *GD/I* varies as $2\,\sigma_G^2/\sigma_{GI}^2$. Jensen's suggestion that *GD/I* be near unity before we suspect bias requires that the interaction evidential of racial bias account for about twice as much variance as does the racial difference which may result from that bias; otherwise the evidence is not deemed to be of practical significance.

Concentration on the statistical representation of group by item interaction obscures the nature of the phenomena being represented. Normally, in interpreting an ANOVA one does not test a fixed main effect if there is a significant interaction. Rather, one concludes that the groups have been affected differentially. For bias Jensen and others go to the other extreme of testing the main effects and ignoring the interaction unless it is extemely large. The logic escapes me. The question being asked of the data is whether there is bias (interaction). We are concerned about bias because it could result in erroneous group differences. The formal or implied index – ratio of group effects to interaction effects – makes large group differences the criterion for lack of bias.

For the central evidence of bias in construct validity – group by item interaction – the proposed criteria would have bias vary inversely with group differences. No statistical analysis will convince those concerned with real bias in real tests that large differences between blacks and whites in test performance are important evidence of lack of bias in those tests.

by **William R. Havender**

Department of Biochemistry, University of California, Berkeley, Calif. 94720

Individual versus collective social justice

Jensen's indisputable conclusions about test bias, or actually the absence of bias, pose a problem only if one believes that parity in group-average statistics is the proper barometer to monitor in making inferences about social justice. And the problem is insoluble. For if one holds that any manifest differences in group "merit" are spurious, then groups currently overrepresented in the economic and educational elites must inevitably be early targets for redistribution, since they cannot have *earned* these exalted positions. This would have only a limited impact on the white gentile majority (since they would lose on average no more than 10 to 20 percent of the positions they now hold), but the impact on such minorities as the Jews and Orientals would be devastating. That Jews have won extraordinary standing is well known, but it is not well known that the Asians (Chinese and Japanese) have, for example, won election to the National Academy of Sciences in a ratio ten times their proportion in the general population. Regardles of

intent, the effect of this social view will assuredly be anti-Semitic. And the flip side of this belief is the eager assent to the pernicious proposition that if a group average difference ever *were* one day demonstrated in an incontestable manner, then the entire structure of liberal democracy would come tumbling down, and we would have no further defenses against reinstituting racially segregated schools and a South Africa-like allocation of jobs on ethnic, racial, and religious grounds. But this is nonsense.

The problem vanishes, however, as soon as one understands that group averages are not the proper statistic to watch in judging the degree of a pluralistic society's justice. And there is an elementary reason why one *should* understand this. For groups consist only of the individuals that compose them, and have no moral standing apart from that of their component persons. This means that *whatever* group ratios result from the just treatment of *individuals* must be a fair one. Readers fortunate enough to be acquainted with Hayek's writings (e.g., 1955) will recognize here an instance of the distinction Hayek draws between "methodological individualism" and "methodological collectivism." Readers not so graced will still recognize this as no more than the familiar philosophy of merit that happily, though precariously, is still officially espoused in this land.

An extremely significant feature of the current attack on standardized testing is the lack of any demonstrably better alternative for making merit allocations. This is the salient distinction between the current debate and its predecessor where, you remember, there did exist a means of readily showing the superior academic capabilities of substantial numbers of people being excluded by the then-prevailing methods. Were this the case today, the remedy would be simple, in fact the same as before: adopt the alternative means as the standard. But no method that has been demonstrated to be superior in detecting hidden academic abilities exists, and the sole animating force now is the disparity in group means.

Nor are the critics of tests willing to abide by the outcome of any conceivable practical demonstration. In this past decade of affirmative action, it has happened many times that large numbers of students have been admitted in place of others who did satisfy the normal criteria, with disproportionate amounts of teaching and tutorial resources expended to nurture any faintest flicker of scholastic ability. The outcome has typically been that no founts of acheivement were uncovered beyond what was already predicted by the usual means. The reaction of these critics, however, has not been to pause and perhaps think, but to vilify those courageous enough to point out the factual nudity of their hypothesis.

The characteristic feature of this debate is the test critics' compulsive focus on this-here mole, and that-there freckle, and that other anecdotal deviation-from-perfection in the application of standardized tests, oddly contrasted with a vast, yawning silence concerning the feasible alternatives, and an unbudgeable unconcern with the outcome when alternatives *are* tried out. This is all too familiar from other debates of the past two decades, notably, the one over Vietnam. If you have an historical memory that extends as far back as a decade, you will recall that we were supposed to focus our critical attention only on the petty bribery of the Thieu regime and napalmed babies, never for an instant on the possibility that the only feasible alternative would be a horror of catastrophic proportions. That horror now confronts us, and "unbudgeable unconcern" is certainly an apt, if too mild, description of most of those who worked to bring it about. One simply cannot refrain from noting this parallel in assessing the balance of credibility in the current flap over mental tests.

As in previous debate, there are clearly deeper motivations at work. For one, there is the manifest intent not to strengthen the use of demonstrated merit in allocating academic opportunities but to *replace* it by political allocation (which means, of course, allocation by intimidation). For another, there is that exaggerated sensitivity, insectlike in that the faintest whiffs of certain scents set the antennae to quivering and the glands to secreting, to any hint that there might be some measure of justice in the current system.

This is a sorry tale. And it is a dreadful sign of the times that those who *favor* individual assessment without regard to one's ethnicity, religion, or race are now forced to argue for the reality of group differences, while those who abominate the distribution of rewards in proportion to proven individual merit and shamelessly prefer making one's chances conditional upon ethnicity have managed to gull much of the media into granting them the halo of virtue.

It must not be forgotten that the unwillingness to permit group properties to fluctuate as the *dependent* outcome of individually fair treatment must have the unfair treatment of individuals as its necessary concomitant. This is an old fight, one that clearly will have to be fought and won again.

by **Jerry Hirsch, *Mark Beeman, and **Timothy P. Tully**

*Departments of Psychology and of Ecology, Ethology and Evolution, and Institutional Race Program, University of Illinois at Champaign-Urbana, Champaign, Ill. 61820; *Department of Sociology and Institutuional Racism Program, University of Illinois at Champaign-Urbana, Champaign, Ill. 61820; **Interdisciplinary Program of Genetics and Institutional Racism Program, University of Illinois at Champaign-Urbana, Champaign, Ill. 61820*

Compensatory education has succeeded

We are presenting specialized commentary on certain aspects of Jensen's overly ambitious attempt and give references which provide documentation and explanation. Though there may be 900 references in the bibliography, in chapter 1, mostly without attribution, Jensen presents list after list of unreferenced criticisms of tests, easily 100, which are then dismissed as uninformed. The unsatisfactory nature of this strategy, which prevents readers from consulting the critics directly, can be appreciated in the case of Professor Banesh Hoffmann's influential critique, because in this instance Jensen (on p. 6) does cite Dunnette (1963 [actually 1964, p. 65]) for "the most detailed and trenchant criticism of Hoffmann's argument . . . using verbal analysis to . . . reject empirical results." Direct consultation of this reference, however, reveals that, on his very next page, Dunnette had conceded: "I would say that empirical validity should not necessarily carry the day over content validity . . . I am in essential agreement with Hoffmann" (p. 66). Hirsch (1976) has previously documented the paramount importance of direct scholarly scrutiny of primary sources, and is still waiting for Jensen to respond; maybe now he will.

After reassuring readers at the outset that he will *not* discuss "the so-called nature-nurture question" (p. xi), Jensen goes on to assert: "We have a theory of intelligne – the polygenic theory – that is entirely independent of any test of intelligence . . . which is now generally accepted by geneticists . . . [with] general agreement . . . that the heritability of intelligence is substantial" (pp. 79, 80, 244). As a valuable antidote to such misleading claims, read the article by the distinguished quantitative geneticist and chairman for 1980 of the Statistics Section of the American Association for the Advancement of Science, Professor Oscar Kempthorne (1978), *commissioned* by the Biometrics Society with specific reference to Jensen, Shockley, and Eysenck. Professor Kempthorne's analysis shows clearly "why the whole IQ-heredity argument as advanced by the hereditarians is deeply unjustifiable and strongly misleading" (p. 21), also see (Goldberger 1979; Hirsch 1970, 1976; Hirsch & Vetta 1978; Hirsch, McGuire, & Vetta 1980; McGuire & Hirsch 1977; Weizmann 1971). Likewise, to appreciate both the unreliability of the references there and how misleading are Jensen's remarks about genetics and "80 percent or more of the IQ variance" in note 3, p. 58, compare its text with Professor I. Richard Savage's (1975) review of his Loehlin, Lindzey, and Spuhler reference. Later, misconstruing heritability, Jensen even claims that *individual* "genotypic value can . . . be estimated" from a kind of regression equation (no. 6.10, p. 243). This is a thoroughly inappropriate application of a population parameter to the individual.

Fundamental to Jensen is Spearman's *g*: "We identify intelligence with *g*" (p. 244). It appears throughout. Witness Spearman's thirty-eight Index entries, at least twice anyone else's, except Jensen's forty. Hirsch has documented Spearman's candor about the rationale for, and his own motivation for, retaining a unitary *g* despite its genetically counterfactual status; in Spearman's words: "the eugenists would be

seriously hindered. Their efforts to better the race could be of slight avail, if they had to be dissipated in hunting after innumerable independent abilities" (Spearman 1914; McGuire & Hirsch 1977, p. 63, see note 2). Except as a single locus or nonchiasma-forming chromosome, which even Jensen does not claim, the mosaic of each genotype fractionates meristically at meiosis and transmits to progeny only a quasi-randomly selected 50 percent (Hirsch 1963); the mythical *g*, therefore, has no biological unity. In the absence of random mating, however, traits with independent genetic correlates can show fortuitous correlations indefinitely (like the evidence for *g*, Hirsch 1967). The mathematics of this relation have been explicated in Li (1955), which source, as Vetta (1977a) has documented, Jensen (1967) has previously misconstrued and has yet to acknowledge (Hirsch, McGuire & Vetta 1980, p. 227).

In addition, despite the widespread recognition that it is scientifically inappropriate and socially misleading, Jensen clings to human heritability estimation because it is equally fundamental to his case: "The substantial heritability of . . . *g*-loaded tests is proof of a biological basis for individual differences in *g* . . . " (p. 251). Furthermore, "*g* is a concept with relevance . . . to understanding species differences in . . . adaptive behavioral capacity." (pp. 250–251).

The typological discussion of "Animal Intelligence" reveals Jensen's gross misunderstanding of "the field of zoology" and the concepts "ethologists" use in our approach to evolution. With respect to contemporary species, despite his misinformed statements about "evolutionary status . . . lower to higher . . . phyletic scale . . . phylogenetic hierarchy . . . phylogenetic levels . . . phylogenetic status" (pp. 175–182), the *scala naturae* was abandoned long ago as a misconception. Read Hodos and Campbell (1969, p. 348), especially their discussions of (1) Jensen's Bitterman reference, (2) Professor Ernst Mayr's admonition: "When the . . . psychologist speaks of The Rat or The Monkey, or the racist speaks of The Negro, this is typological thinking," and (3) "the implication that the particular species being investigated is a generalized representative of the entire order or class when in fact that species may be highly specialized and not at all representative." Among the 1,000,000 or more species of animals, the species count per group exceeds 200 for primates, 3,000 for rodents, 8,000 for birds, 30,000 for fishes (News and Views 1980), and so on, most of which have *not* been studied behaviorally. Within any group, among those that have been studied, the differences are striking. (See also the excellent review of Jensen [1980] by Gould [1980] which reached us during the typing of our text.)

Readers interested in a truly empirical, experimental approach to animal intelligence and to the question of test bias should consult Harrington (1975; or McGuire & Hirsch 1977, p. 33), who has built on the foundation laid by Hebb. Harrington has shown in experiments (see Kempthorne 1977, 1978 on the fundamental distinction between experimental and observational studies with respect to the inferences that can be made) that, when the proportions are varied for different races present in the racially mixed populations on which tests are standardized by routine psychometric procedures and the resulting tests are then administered to the separate races, there is a strong correlation between the level of representation of a race in the standardizing population and the success of that race on the test. The higher the proportionate representation of a race in a population, the better that race scores on the test standardized on that population and vice versa: "generalization from these data to man is direct and not analogical: the experiment was an empirical test of common psychometric assumptions and procedures. Generalization is therefore to those assumptions and procedures. The implications are far ranging. Majorities will score higher than minorities as a general artifact of test-construction procedures" (Harrington 1975, p. 709). We hope that in his response Jensen will discuss this important work in greater detail than space allows us.

Tests interact not only with genetic groups but also with the presence or absence of benign environmental conditions. To help readers appreciate the human condition, we challenge Jensen to contrast in his reply his eugenic approach with the well-documented environmental interventionist evidence reported by Darlington et al. (1980; also see Sewall & Howard 1979) – *compensatory education has succeeded.*

Note

As I have shown previously, (Hirsch 1976), "He stumbles repeatedly in biology . . ."; now he [Jensen] reports cholinesterase to be a neurochemical transmitter. It is not. In fact, it is literally the opposite, being an enzyme that destroys the transmitter!

***by* Lloyd G. Humphreys**

Department of Psychology, University of Illinois at Urbana, Urbana, Ill. 61801

Intelligence testing: the importance of a difference should be evaluated independently of its causes

It is easy to misinterpret the message in this book, but it is important that the message be interpreted correctly. Jensen does a bit to facilitate the former and to make more difficult the latter, but many of his readers will do a good deal more to confuse matters. This reviewer will try to clarify the issues and help readers achieve a balanced point of view concerning the significance of the massive amounts of data presented.

The author states at the outset that he intends to avoid the heredity-environment debate with respect to race differences in intelligence. This is laudable, and his intent should be taken literally by the reader. A careless reader misses, however, the significance of the author's definition of intelligence and keeps thinking about the issues in terms of the definition of the term common to our culture. That is, the habit of thinking about intelligence as a fixed capapcity of the organism is deeply ingrained. Intelligence is defined by Jensen, however, as a phenotypic trait like height and weight. It may not be as easily observable or measurable as traits of physique, but it is basically similar.

Defining intelligence as a phenotypic trait is necessary and not done to confuse. It is essential that man-in-the-street thinking about intelligence be abandoned. Jensen does contribute to the unfortunate mental set of most readers, which is compounded by their fixed attitudes toward him, by referring several times to the supposedly well-established high heritability within groups of the phenotypic trait. It is all too easy for the reader to generalize from high heritability within groups to high heritability between groups. Given Jensen's intentions, that is, to concentrate on bias in the phenotypic measure, one might have expected an explicit warning against such a generalization. One might also have expected some recognition of the fact that other authors who are not environmental ideologues have made lower within-group heritability estimates than the ones he presents.

Jensen's definition of bias will also escape many readers. Based on the several classes of data presented, it is also entirely on the phenotypic level. These data do not refute either genetic or environmental causation. For many readers a test is biased if it does not measure *real* intelligence. Whatever that may mean, it is clear that it is far removed from a comparison of slopes and intercepts of regression lines or the size of an interaction between race and item difficulties. The debate concerning intelligence has been so completely focused on the nature-nurture issue, and has been so emotional, that it is difficult to view the matter from any other perspective.

If we accept, as I think we must, that neither the largely genetic nor the largely environmental hypothesis concerning racial differences in intelligence can be rejected by data meeting acceptable scientific standards, there is still one firm conclusion that can be reached from the data Jensen presents. 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 predominantly genetic or environmental.

There are two incorrect conclusions that are likely to be drawn from the evidence that Jensen has marshalled. One is to dismiss the environmental hypothesis is on the grounds that the evidence appears to refute it compellingly. The second is to dismiss the evidence and to belittle the importance of the difference. With respect to human traits, the heredity-environment causal dimension is relatively independent of the dimension of importance. Color and taste blindness are genetically determined, but are not terribly important in human affairs. Being able to speak a foreign language without accent is environmentally determined, but is strongly resistant to change and is at least as important as color vision. It is possible to conclude that the race difference is important without drawing any causal inference whatsoever.

Is there something analogous to the acquisition of a foreign language that takes place in intellectual development? I do not know, but I do contend that we know all too little about early intellectual development in the human to reject that notion out of hand. If there are similar mechanisms, it is clear that environmental intervention should occur early and that even young adulthood is many years late. There are also bits and pieces of information, adding up to hints and promises, that the notion is plausible.

The reader who seizes on bits and pieces of information to explain away the phenotypic differences in intelligence between blacks and whites is missing the point. An environmental explanation, no matter where it stands on the continuum between wishful thinking and sound scientific documentation, does not reduce the importance of the deficit. Plausible environmental explanations are worthless unless we can devise and apply effective interventions. The possibility of doing something effective depends directly on recognition of the seriousness of the problem and on accepting its dimensions as revealed by the research reported in this volume. If we accept these facts, perhaps we can use the hints and promises to devise programs that stand more chance of being successful in closing the gap than affirmative-action programs at age eighteen and beyond.

***by* Oscar Kempthorne and Leroy Wolins**

Department of Statistics, Iowa State University, Ames, Iowa 50011; Departments of Psychology and Statistics, Iowa State University, Ames, Iowa 50011

Controversies surrounding mental testing*

Each of us is in our own way accountable to society, and in every society testing is widely used in many forms to evaluate accountability. There can be no doubt that standardized tests are necessary, and, despite their imperfections, represent a vast improvement over arbitrary and unfairly administered evaluation procedures of the past. Of course, there must be a rational realization of *what* is being done, and this is where the problem lies. Do these tests provide fair bases for evaluating individuals?

Arthur Jensen has made a huge effort to answer this question. Doing justice to this effort demands that we should *not* impute Jensen any motivation (e.g., racial bias) that is not evidenced in this book.

The central theme of this book is that intelligence tests provide fair means of assessing the intelligence of native-born English-speaking Americans. The author attempts to support his claim by educating a nontechnical audience in principles and uses of psychological measurement. The difficulty of the content is variable. Parts are very complex, outstripping even the author's technical knowledge and acumen, as evidenced by many mistakes and inconsistencies. Jensen seems to perceive statistical methods as a mechanical tool which the researcher uses to pry information from data. In fact, these statistical procedures are *based* on models and, thereby, are dependent on the truth of the models.

Despite all the mistakes, this is an important book. The central theme comes out clearly. The basic issue is psychological tests – not race differences. What is at stake is years of careful research and development of psychometric devices. Since the topic is bias in testing, one cannot ignore black-white differences, since most research on test bias involves these two groups.

Large sample results on societally relevant criteria invariably indicate that blacks perform less well, on the average, than whites. Blacks perform less well than whites on standardized tests, on the average, probably for the same unknown reasons.

Some of us would have regarded it as fortunate if the tests had demonstrated that black-white performance differences on tests did not exist, so that we could attribute the associated differences on outside, societally relevant criteria to prejudice. This did not occur, despite arduous efforts to develop tests that were "unbiased." Research indicates that tests that do not separate whites from blacks are not valid predictors of societally relevant criteria for either whites or blacks, whereas tests that turn out to separate whites from blacks are valid for both whites and blacks.

Adversaries of psychological testing should recognize that tests do not cause race differences, and banning the tests will not solve the problem of race differences on societally relevant criteria. Fostering psychometric research may help in understanding the causes of these race differences and contribute to the solution of the problem. We may fear this understanding, but to seek it is the rational course.

Much of Jensen's book appears excellent. It contains many examples of thoroughly reasoned and well-documented conclusions, but it cannot be regarded as an authoritative source because of pervasive conceptual and methodological mistakes. We recommend this book for those with strong psychometric and measurement backgrounds. We do so because, despite shortcomings, we believe it makes a good case for its central theme and contains many relevant ideas and research results.

Some specific commentary follows, predicated, partially, on Jensen's intended audience.

Not all of the criticisms and questioning presented in chapter 1 are invalid, as Jensen seems to imply. Societal procedures that affect lives and opportunities of individuals must be questioned, especially by the adversely affected minorities. Also, we seem to be told that only if you are a psychometrician can you understand the issues and techniques. We reject this. Mental testing should be under fire, which is not to mean that it should be "canned." The testing industry is to blame for not taking criticisms and questions seriously enough.

In chapter 2, landmark court cases are reviewed and discussed, providing a reasonable and interesting backdrop for the whole book, but some of the judicial remarks suggest that we cannot rely on our court system for reasonable judgments. What is the phrase "innate learning abilies" (p. 29) supposed to mean? *We suggest strongly that that phrase be banned from the professional literature.* The fact that 12th-grade performance on a test is predictable from 6th-grade performance does *not* imply these performances are "innate."

In chapter 3 we find that "discrimination" is always present and necessary for many societal purposes and that no test has ever been constructed for the purpose of racial or social discrimination. We agree, but disagree with Jensen's efforts to "apportion the total [IQ] variation" into a number of "sources." He finds "that race and SES [socioeconomic status] contribute only 22 percent to the total IQ variation." This analysis is either useless or it indicates that it is not the case that "relatively little" of the total variation is associated with race and SES. The graph of Full Scale IQ against SES (p. 44) is an accurate reporting of hard facts and it implies that if blacks have a legitimate case against testing, then so do whites of low SES. That "the use of objective tests . . . has promoted social justice" (p. 57) is clear, but the question not adequately addressed is whether such use has also promoted social injustice.

In chapter 4 Jensen is unsuccessful in telling us why testers have "settled on" the Gaussian distribution. The discussion of interval scaling is untenable. Our reaction to Jensen's discussion of the heterogeneous content of IQ tests is that it excludes musical, artistic, and cooking abilities, even though these abilities also contribute to the "good society." We must reject the idea that we can map the whole range of abilities on a single number line, but to postulate an infinity of abilities is futile. The beginnings of any science are classification and prediction, and tests are constructed to predict success within a wide

variety of vocations. Test constructers are not to be classified as gnomes who are foisting their own prejudices on the outside world. In this chapter, pointless obscurities and associated controversies are called forth with the statement, "The polygenic theory of individual variation in mental ability leads us to expect a more or less normal distribution of ability."

Chapter 5 is a reasonable exposition of "Varieties of Mental Test Items" but it is not the sort of exposition we are entitled to hope for. The items on pages 148–150 involve both *actual knowledge of the language and ability to use language properly.* How then can one use such items to "show" that children who do not know the language have low mental ability? It is surely justifiable to use such items to develop a test of language comprehension. Without adequate discussion, the atrocities of the early part of this country (in which individuals of, say, Russian origin were labelled as morons) seem justified. If a test battery of such items is used to tell parents that their children have so little familiarity with the language use of ordinary schools that they need special treatment, then it is hard to see how rational criticism can be mounted. Jensen appreciates this, but why is there no discussion?

In chapter 6 Jensen asks, "Do IQ Tests Really Measure Intelligence?" At one time Jensen said, we believe, "intelligence is what intelligence tests measure." Here, Jensen quotes the reasonable homily of Wechsler on page 171. But it leaves us with much uncertainty. The statement of Wechsler does not tell us what [intelligence] tests measure. We see, rather frequently, the question "Is Intelligence a 'Thing?' " (c.f. Gould 1980). What is the meaning of the question? We suggest that this is not a well-posed question. Let us ask, in the same temper: is temperature a "thing?" Intelligence is a construct. It is not a "thing," and no one should believe it is.

When we get to "Armchair Analysis Versus Empirical Investigation," we are given intelligence A, B, and C. The presentation and discussion is highly defective. Jensen brings in genetics but seems to be unaware of the logical and statistical difficulties that ensue.

We get to the correlation matrix and factor analysis, starting with *The General Factor.* The exposition is "not bad," but there are great logical difficulties that are properly discussed in Mulaik (1972, esp. pp. 133, 135, 173 and 327).

We find that "No really clear distinction can be made operationally at the level of tests between *intelligence* and intellectual achievement" (p. 250). If only this were on the front burner of Jensen's mind rather than on the back burner, we would not have all the controversy associated with Jensen's writings. A real problem in this whole book is, we suggest, that Jensen rides every horse at one or another place, and does not seem to realize that the horses are going in every possible direction. Technical errors are too numerous to discuss.

Jensen's general discussion of reliability in chapter 7 seems reasonable and adequate for the purposes of his book. The large tables of reliabilities have considerable force as *experiential* facts. But Jensen is in trouble when he gets to the factor analysis of the intercorrelations of Binet IQs at various ages. The factor analysis that Jensen does is conceptually invalid. This chapter contains many other statistical and conceptual errors.

Chapter 8 is where the real action lies. Jensen enumerates and discusses the four C's: content, criterion, concurrent, and construct validity. We found the discussion of construct validity not entirely satisfying. But we read: "IQ has more behavioral correlates than any other psychological measurement" (p. 313). There seems to be no way of combating this, since it is an experiential fact. Finally, "does IQ predict scholastic achievement and then does scholastic achievement predict job performance?" *The experiential fact is that they do.*

In chapter 9 we reach the topic of the title of the book. Although we do *not* find a reasonable general *definition* of test bias here, we agree with Jensen's ideas and procedures for examining test bias. Jensen claims (p. 515): "In the vast majority of studies, the regressions of criterion performance on test scores do not differ for blacks and whites." He presents, we think, an honest account of the situation.

Chapter 10 is well done, and we accept his summary statements (p. 515): "differential validity for the two racial groups is a virtually nonexistent phenomenon" and "in the vast majority of studies, the regressions of criterion performance on test scores do not differ for blacks and whites."

Mental tests are biased if one takes the view that they are to give a constant result regardless of environment and education. But we insist that this is a *completely absurd* view. A process of measurement of body weight is not biased if it gives 90 pounds as the weight of a twenty-one-year-old male of height 6 feet. The process tells us what we need to know, the weight of the man. It does not tell us the causation of this very low weight.

The general message of chapter 11, which we judge to be quite well-supported, is that the correlations between tests do not differ between racial groups. This is a necessary condition for validity of comparisons of means, but it is, of course, not sufficient.

In chapters 12, 13, and 14, "External Sources of Bias," "Sex Bias," and "Culture-reduced Tests and Techniques" are covered. We accept Jensen's view that the first two do not lead to significant problems and a reasonable account of the last item is given.

In chapter 15, "Uses and Abuses of Tests," we find reasonable discussion and some problems in interpreting subpopulational profile differences using normalized scale scores (pp. 729–730). These data suggest real problems with the supposed "culture-fairness" of the tests. We do not agree that "ability grouping at the elementary school level is more a convenience for teachers than a benefit to pupils" (p. 739).

Finally, we are in very strong disagreement with Jensen's past writings on genetics and IQ, and on genetics and education. We have taken the position that Jensen's *other* writings should not enter the evaluation of this book, except to the extent that they are cited and some of their lines of thought are again used. It would be comforting if we could dismiss the present book as easily as we can dismiss Jensen's "hereditarian" writings, some of which he repeats in this book. But we find that we cannot. The role of genetics appears in the present book significantly, but not significantly in relation to the question of bias in mental testing.

This book tells us that our society has a huge problem, which Jensen chose not to address, but the problem is not the bias of tests.

Editorial Note

*A longer review by these authors, containing specific technical criticism, will appear in a forthcoming *BBS* Continuing Commentary on this topic.

***by* Paul Kline**

Department of Psychology, University of Exeter, Exeter EX4 4QG, England

Test bias and problems in cross-cultural testing

A major problem in the testing of minority groups concerns the validity of the testing instrument within such groups. Although this is highlighted in cross-cultural testing, when inhabitants of different cultures are compared, the same difficulties arise in the testing of minority groups within the same culture, as occurs in the U.S.A.

Jensen is aware of this problem, and in chapter 14 attempts to deal with it. His solution is essentially the construction of items with little specific Western cultural loading; for example, pantomime instructions rather than oral, familiar content rather than rare.

Now this entirely ignores the real difficulties of cross-cultural testing as conceived by those who have worked in the field, and unless these problems are faced and discussed, critics of Jensen's approach will be well armed.

Cross-cultural problems in testing abilities. One problem is summarised in the emic/etic dilemma (e.g., Triandis, Malpass, & Davidson 1971). The emic approach argues, in the anthropological tradition of Malinowski and Boas, that cultures have to be understood within their own terms. Behaviors have to be studied as they are perceived by members of that culture. Intelligence and intelligence-testing well exemplify this point. In the West both the concept and the intelligence test have a definite meaning. Wober (1973), however, has shown that intelligence is quite differently conceived in Uganda, and thus, on the emic view, to compare the results of Ugandans and Westerners is not meaningful.

The etic approach, on the other hand, seeks universalities. Jensen is clearly in this tradition (cross-cultural comparisons cannot be emic

by definition). However, the trouble with these comparisons is that they are meaningless. Williams (1975), in the face of this dilemma, argues that the ideal is to develop emic measures of etic constructs. The problem then becomes one of trying to ensure that the emic measures are equivalent. The aged African chief who, on being given the Porteus Mazes test, was asked to imagine it as a cattle kraal to which he should lead his cattle, and refused to try the item on the grounds that anyone who constructed a kraal like that was mad, highlights the point. So too do the Gurkhas tested by Warburton (1951), who could not recognise an apparently lifelike picture of the god Kukri, which they all carry.

Jensen does not raise these points, claiming that they lie beyond the scope of this book, but this merely avoids the difficutly. His solution is to distinguish between predictive validity and construct validity in cross-cultural studies. Thus he argues that if a test predicts in the new group, as it did in the old, then it is satisfactory in the new culture. Construct validity in cross-cultural studies he admits is difficult to demonstrate. This argument, however, will not do. It is analogous to the distinction between criterion-keyed and factor analytic tests. Criterion-keyed tests such as the MMPI may well discriminate groups, but this tells us nothing about the psychological nature of the variable they measure. Thus an ability test may well have predictive validity in a minority group, but it tells us, on this account, little about the abilities of the group. As Irvine found (1969) in his studies of Rhodesian abilities, *g* and the other main Western ability factors could be extracted; yet he argued that there were African abilities largely untapped by the tests. Thus on the basis of Western ability tests, though predictably valid and loading as in the West, not much could be said about African abilities. If this is the case, then there is little hope of understanding, on this basis, cognitive abilities in Africa and their development.

Thus it is with Jensen's empirical approach to the testing of cross-cultural groups. Unless the variables have demonstrated construct validity in the new cultures, little psychological insight will be gained into them.

There is a further inconsistency in Jensen's approach to this problem. He adopts, advisedly, Cattell's (1971) factorial view of intelligence: fluid ability, the constitutionally based mental ability, and crystallised ability, fluid ability as it is envinced in a culture. Hopefully, he advocates for the study of minority groups that we attempt to measure fluid ability, with culture-fair or culture-free tests such as those of Cattell or Raven's matrices. Apart from the practical difficulty of obtaining pure fluid ability measures, it must be realised that part of the Cattell theory claims that fluid ability is invested in crystallised abilty, the skills valued by a culture, the cultural expression of fluid ability. It is crystallised ability as measured by the WISC and the Binet that correlated highly with cultural achievement. Thus, for measuring intelligence one needs measures both of fluid and crystallised ability. The latter, however, is inevitably culture-based and only a proper analysis of cultures will permits its measurement. The mere establishment of predictive validity is not sufficient.

This aspect of Jensen's work seems simplistic. It is a particularly serious defect since it must to some extent undermine any conclusions he draws about national and cultural differences in intelligence. Anthropological psychology cannot be any longer ignored, even though its conclusions may be refuted.

***by* Langdon E. Longstreth**

Department of Psychology, University of Southern California, Los Angeles, Calif. 90007

The definitive work on mental test bias

Jensen's book is the definitive work on mental test bias. It is hard for me to imagine a more thorough, more scholarly, more *objective* treatment of the subject. I emphasize the word objective because Jensen has often been accused of the opposite by his many detractors. Yet it is all here: the data, the reasoning, the conclusions. The data are numerous. The reasoning is a model of clarity. Together, they force strong conclusions. I will mention six that strike me as particularly relevant to the position of Total Egalitarianism espoused so often (i.e., all groups of people are equal in intelligence; ergo, if differences are detected, the instrument is faulty). I will state these conclusions in the negative to make clear the Total Egalitarianism assumptions they deny.

(1) IQ scores are *not* arbitrarily distributed in a Gaussian fashion by test-makers. Rather, the normal distribution of IQ scores reflects a fundamental property of intelligence, in the same fashion that the normal distribution of physical traits such as height and weight reflect fundamental properties of these variables. That there exists a small proportion of the population with superior mental ability is not the result of a man-made elitism constructed to lord it over the rest of humanity, any more than the small proportion of 7-foot basketball players is the result of a special "height" scale so constructed by basketball coaches to lord it over competing teams.

(2) Intelligence is *not* totally, or even mainly, culture-dependent in its definition. Kagan, among others, is wrong to argue that "At another place and time a different set of skills might be primary and the child who was intelligent in terms of the first set might be very unintelligent in terms of the second" (Jensen 1980, p. 247). Such cultural relativism is wrong because it confuses what is "primary" (important for survival) with *g*. The two bear no necessary relation to one another, and an intelligent Bushman is more than a speedy runner with a strong arm, just as an intelligent basketball player is something more than a 7-foot man.

(3) IQ tests are *not* simply achievement-test predictors. Positive manifold, or *g*, cannot be denied, and it exists in the face of IQ tests that deviate far and wide in form and substance from academic achievement tests. IQ scores have more significant relationships with other behavioral characteristics – occupation, job rating, even morality – than any other single psychological index. It is, in other words, an entirely respectable and important scientific concept.

(4) By and large, IQ tests are *not* biased against blacks in the prediction of academic achievement. On the contrary, it is almost uniformly the case that where the test yields less accurate predictions for blacks than for whites, the achievement of blacks is over-predicted. Thus it is more accurate to speak of bias *against whites.* Needless to say, not too many proponents of test bias have added this refinement to their argument.

(5) By and large, black-white differences in IQ scores are *not* the result of "white test items" that bestow an advantage upon any person raised in that culture. The absence of race-by-item interactions that account for a sizeable portion of the variance in test scores rules out such a notion. Blacks tend to perform more poorly on *all* test items, and on *all* IQ tests, regardless of where they are administered, by whom, when, and so on. In fact, the items that best discriminate between one black and another black tend to be the same items that best discriminate between blacks and whites.

I cannot resist an aside here. I have frequently run across the argument that a given IQ test is biased against this minority or that because the minority was not included in the standardization sample. This argument has always puzzled me because I have never seen its logic made explicit. Jensen is the first person I am aware of who challenges this notion, calling it the Standardization Fallacy. I hope his challenge stimulates those who subscribe to the notion to think it through, and to publish their defense – if there is one.

(6) Piaget-type tests do *not* eliminate or reduce socio-economic status (SES) and race differences in test scores, and they *do* correlate with standard IQ-test scores. Piaget's items have been heralded as superior indices of intelligence because the answers to them are presumably learnable in all cultures, in all races, as a result of day-to-day experiences common to all human beings of a given age. It is not generally acknowledged that such items yield the same kinds of relationships to other variables as do the regular IQ-test items.

These conclusions, as well as a host of others, make it clear that test-bias proponents have their work cut out for them if they wish to talk about the real world. And yet these conclusions, along with their supporting arguments and data, are really not the heart of the book at all. The heart of the book, at least to me, is chapter 9: "Definitions and

Criteria of Test Bias." This is a strictly methodological chapter, and its ninety-six pages provide almost a step-by-step description of the various ways in which predictive and internal test bias may be measured. There are many lessons to be learned here, not the least of which is how various item statistics are interrelated and how they are related to the analysis of variance with groups and items as the independent variables. This is a gold mine of information, providing indispensable information for the serious student of test bias.

There is one issue Jensen did not address, and I hope he will discuss it in his reply to these comments. It has been argued that standardized IQ tests might predict standardized achievement test scores, but not other indices of actual classroom performance, and this is especially the case for minority and low-SES students. So common-methods-variance inflates the observed IQ-achievement correlation. Evidence: when a nontest index is used, such as teachers' grades, the IQ-grade correlation is trivial. Conclusion: do not use standardized IQ tests at all.

In a study report too late for inclusion in Jensen's book, this position is seriously challenged (Messé et al. 1979). It is strongly implied here that it is teachers' grades that should be thrown out if anything is to be discarded owing to lack of validity. A large-scale study of British schoolchildren is then described (N = 5,200) in which special efforts were made to improve the validity of teachers' ratings. Result: standardized ability scores correlated .60 with teacher ratings, and were not affected by SES level of students. Regression intercept differences were found, however, reflecting bias *against* upper-SES children: their teacher ratings were underpredicted, while low-SES children's ratings were overpredicted.

by **R. Travis Osborne**

Psychology Department, University of Georgia, Athens, Ga. 30602

The Spearman-Jensen hypothesis

Test question: How would the man who gave us *How Much Can We Boost IQ and Scholastic Achievement?* (Jensen 1969) and, courtesy of a New York newspaper, the term "Jensenism" reply to the following questions. Do you:

(a) advise routine IQ testing of school children?

(b) recommend minimum competency testing (MCT) for school graduation?

(c) advise ability grouping, homogeneous grouping, or tracking?

If you believe he would have replied in the affirmative to any of the three questions, you should read chapter 15 of *Bias in Mental Testing (BIMT).* Other readers are encouraged first to examine the Preface where they will learn that *BIMT* is not an "easy" book or a showy textbook. The author's main purpose is to examine every aspect of bias in mental testing, from ability grouping to z-scores. Arthur Jensen is no newcomer to this field, as some of his critics charged he was in genetics after he published *Genetics and Education* (Jensen 1973). His early graduate research was directed by Kenneth Eells who, over thirty years ago, along with Allison Davis, Robert Havighurst, Virgil Herrick, and Ralph Tyler, conducted the first comprehensive investigation of intelligence and cultural differences.

Jensen's latest book, by far his best, is really two books in one. The first eight chapters would make an authentic text for psychometric students in education and psychology. Law students and journalism majors would find these chapters profitable.

The last half of *BIMT* is more suitable for advanced graduate students interested in mental test theory. This is the "hard" part. Here, on occasion, the reviewer was tempted to invoke the frankness of E. L. Thorndike (1905), who said, "I take Mr. Spearman's method of correction for attenuation on trust as I do not possess the mathematical knowledge to derive his formulae." Except for the nontechnical definition on page 48, no other mention is made of bias until chapter 9, page 367, where are given precise definitions of bias and the rationale for various criteria and methods of statistical detection of test bias. Bias is not to be confused with unfairness, a philosophical position based on the fair use of tests.

Construct validity criteria for test bias are complex but methods given by Jensen permit the evaluation of various hypotheses of cultural bias. If a test behaves the same way for different groups with respect to a number of features of test performance, the test is presumed to be unbiased for those groups.

After guiding the reader through the methodological rocks and shoals, Jensen brings together in the final chapters massive evidence from primary sources that demonstrates convincingly "that most standard ability and aptitude tests in current use in education, in the armed forces, and in employment selection are not biased for blacks or whites with respect to criterion validity and that the little bias that has been found in some studies has been in a direction that actually favors the selection of blacks when the selection procedure is color blind." His detractors will, of course, interpret Jensen's omissions of heritability of IQ from *BIMT* as a hasty retreat from his earlier strong hereditarian position. Not so. Jensen says "even an elementary explication of heritability is beyond the scope of this book."

To keep his critics alert, no doubt, Jensen does report, almost as an aside, that myopia is believed to be attributed to genetic factors and myopia is quite markedly associated with higher IQ. No purely environmental explanation has been found, nor can this reviewer imagine one. Karlsson (1978) concludes that "the myopia gene has an important stimulant effect on brain activity. It thus becomes the first identified specific gene which appears to contribute significantly to intelligence."

During the last twelve years, environmentalists seem to have been so obsessed with the trees that they have overlooked the forest. They chortle over Jensen's typos and other insignificant errors while failing to come to grips with the hard core on which Jensen stakes his claim. The Spearman hypothesis has been the bedrock of all of Jensen's important work since 1969. Simply stated, it says that almost any and every test involving any kind of complex mental activity correlates positively with any other test including complex mental activity. All such tests measure a common factor to some degree, which accounts for the intercorrelations among all the tests. Spearman called this common factor "general intelligence," or simply *g*. To Jensen, one of the most important findings of cross-racial factor analytic studies of a variety of cognitive tests is their complete consistency with the hypothesis originally advanced by Spearman, that the magnitude of the black-white difference (expressed in standard deviation units) is directly related to the test's *g* loading.

To discredit the Spearman hypothesis with replicable hard data, something environmentalists find in short supply, would at the same time dispose of the hard core of Jensen's research and cripple his own hypothesis of Level I and Level II abilities. In this latest book, Jensen's challenge is loud and clear.

Some say the environmentalist's research program is degenerating. Peter Urbach (1974) sees some hope because resourceful environmentalists of the future may well invent a powerful heuristic which will lead them to the explanation of individual and group differences in IQ.

Important new findings, released perhaps since *BIMT* went to press, further confirm the Spearman-Jensen hypothesis. Until the Congressional Hearing on May 15, 1979, both ETS (Educational Testing Service) and CEEB (College Entrance Examination Board) had successfully avoided the race issue in their widely distributed research reports. In a lengthy report presented to the Committee, W. H. Manning, Sr., Vice-President of ETS, said: "The results have been quite consistent. Differences in test score averages across ethnic groups are consistent with actual performance in college. All other indications in these studies point to the conclusion that tests typically predict the same way with the same validity for whites and minorities" (ETS 1979). The report goes on to name the investigators and data sources.

Bias in Mental Testing deserves a place in the professional's library alongside the several *MMY (Mental Measurements Yearbook)* volumes. The graduate student or investigator planning research, an attorney writing a brief, or a judge hearing a case needs the information of *BIMT*. Had the WISC data in *BIMT* been available to Judge Peckham (1979) in the case of Larry P., the judge would not have been misled concerning the bias of WISC items.

by **Cecil R. Reynolds**

Buros Institute of Mental Measurements, University of Nebraska-Lincoln, Lincoln, Nebr. 68588

In support of *Bias in Mental Testing* and scientific inquiry

Jensen has provided an in-depth analysis of the single most crucial hypothesis facing scientific and applied psychology today. The cultural test bias hypothesis contends that all group differences in mental test scores are due to a built-in cultural bias of the tests themselves; that is, group score differences are an artifact of current psychometric methodology (Harrington 1975). If the cultural test bias hypothesis is ultimately shown to be correct, then the 100 years or so of psychological research in human differences (or differential psychology, the scientific discipline underlying all applied areas of human psychology including clinical, couseling, school, and industrial) must be dismissed as confounded, contaminated, or otherwise artifactual. Psychology, to continue its existence as a scientific discipline, must confront the cultural test bias hypothesis from the solid foundations of data and theory and not allow the resolution of this issue to occur solely within (and be determined by) the political *Zeitgeist* of the times. *Bias in Mental Testing* is a strong step in the right direction.

Jensen's new book provides a thorough and dispassionate review of virtually all empirical research relevant to the evaluation of cultural bias in psychological and educational tests that was available at the time the manuscript was prepared. Over the last eighteen to twenty-four months, however, a substantial body of literature has become available regarding cultural bias in the psychological assessment of children (an area Jensen notes as being rather meagerly researched compared to the large number of studies with adults). These studies have been reviewed in detail elsewhere (Reynolds, in press a), and will not be referred to specifically here. The results of these studies of bias in assessment with children provide even stronger support for Jensen's conclusions that differential and single group validity cannot be substantiated at all by current empirical evidence.

Consistent with Jensen's reports of studies with adults, analyses of content validity of children's aptitude tests, using standard ANOVA (analysis of variance) methodology to examine for a significant group by item interaction, typically find that only 2–5 percent of the variance in performance on such tests is due to biased items. Even this small impact on scores is believed to be a spurious, methodological artifact (Hunter 1975), since few aptitude tests are entirely unidimensional. Basically, the same results have been reported for achievement tests, although the use of the ANOVA methodology is questionable in examining for bias in achievement test items whenever more than one classroom, or especially more than one school or school district, is employed unless there is nearly perfect, proportionate representation of all groups in all classrooms. More convincing, and more important, are the many studies of bias in construct validity of intelligence tests that have recently become available.

A number of studies are now available reporting factor analyses of the Wechsler Intelligence Scale for Children – Revised (WISC-R), the most widely used individual intelligence test for children, for large groups of black and Mexican-American children. When the results of each of these studies are compared to results for white children only from the WISC-R standardization sample, an amazingly consistent pattern of similarity occurs. The median coefficient of congruence for the *g* factor, two-factor solutions, and three-factor solutions ranges from .91 to .99 with a median value of .96. Factorial invariance has been noted to occur across race for children irrespective of whether comparisons are based on correlation matrices or, in a more rigorous methodology, on covariance matrices. New research is also available to indicate that Jensen's conclusion that the correlations of raw scores with age on aptitude tests are constant across race and sex (see chapter 11) is generalizable to many other types of specific aptitude scales and highly *g*-loaded tasks (Reynolds, in press b). In the examination of slopes of the regression between age and raw scores in the above study, a general trend for the scores of black males to increase with age at a lesser rate than those of females and whites on more highly *g*-loaded tasks was also noted, again a finding consistent with Jensen's conclusions.

Many studies of bias in the predictive validity (systematic or constant over- or underprediction of criterion performance as a function of group membership) of IQ tests for minority and nonminority children from the ages of 5–17 have also recently been completed. In a variety of intelligence and aptitude tests with numerous academic achievement variables as criteria, the only bias found favors minority group members. That is, in all cases where bias occurs, minority performance on the criterion has been overpredicted (these studies are reviewed at length in Reynolds, in press a). The disproportionate number of minority children in special education programs cannot be accounted for on the basis of bias in psychological tests but rather, when referrals for gifted programming are eliminated, by the substantially higher referral and failure rates of these children when in regular education-classrooms.

In spite of the large body of evidence reviewed by Jensen and the continued support of his conclusions by the continuing empirical research on bias, the issue of bias in assessment cannot (nor should it be) laid to rest. The controversy over bias in assessment will remain with psychology perhaps as long as the nature/nurture controversy, and perhaps even for similar emotional reasons. The issue is crucial enough that investigation must continue with new tests, new criteria, new methodologies, and even new paradigms from which to view the empirical questions. Jensen's volume makes an important statement in challenging a socially, politically, and emotionally charged scientific issue and meeting it head-on with empirical research. Psychology and its practitioners must continue to assault these and other controversial issues within the domain of rational scientific inquiry, lest psychology be reduced to an impotent science whose major issues are resolved not in the scholarly court of research, theory, and inquiry but in the judicial courts of the land, as already attempted in California (Peckam 1979).

by **Robert Rosenthal**

Department of Psychology and Social Relations, Harvard University, Cambridge, Mass. 02138

Error and bias in the selection of data

This commentary allows us to bring into juxtaposition the present BBS multiple review of Arthur Jensen's book with an earlier BBS target article by Rosenthal and Rubin (1978). Jensen discusses in some detail the results of studies of interpersonal expectancy effects (pp. 607–609 of *Bias in Mental Testing*). The purpose of this commentary is to correct errors of fact and to point out the considerable bias operating in the target volume's selection of evidence for presentation.

The Pygmalion experiment (Rosenthal & Jacobson 1968) was described as failing to stand up under critical scrutiny. No mention was made, however, of the fact that in the most ambitious of the critiques of Pygmalion, that by Elashoff and Snow (1971), the results were indistinguishable from those reported by Rosenthal and Jacobson both with respect to significance level and with respect to effect size (Rosenthal & Rubin 1971). For eight transformations of the Pygmalion data made by Elashoff and Snow, *every* one reached significance when significance had been claimed by Rosenthal and Jacobson (Rosenthal & Rubin 1971, p. 141).

In the target volume an experiment by Seaver (1973) was admitted to be in support of the hypothesis of interpersonal expectancy effects. However, it was not viewed as in support of the Pygmalion hypothesis because it employed as a dependent variable an achievement test rather than an IQ test. Yet three different studies showing no effect of teacher expectations were counted as evidence *against* the Pygmalion hypothesis, although they too employed only achievement tests as dependent variables rather than IQ tests (Dusek & O'Connell 1973; Gozali & Meyen 1970; Pitt 1956). In short, when achievement test results favor the hypothesis they are excluded from evidence bearing on the Pygmalion effect. When they go against the hypothesis they are included as evidence bearing on the Pygmalion effect.

Also singled out for comment is a study by Deitz and Purkey (1969)

Table 1 (Rosenthal). *Summary statistics for studies selected by Jensen and for other subsets of studies*

	Number of studies	Mean effect size	Mean *Z*	Proportion of studies reaching $p < .05$
Jensen's studies	13	.00	.04	.00
All dissertations with special controls*	18	.78	1.86	.56
All studies with special controls*	43	.64	1.70	.56
All studies of everyday situations	112	.88	1.03	.40
All studies	345	.70	1.22	.36

*For both cheating and observer errors.

"as it revealed no expectancy effect based on pupil's race" (p. 608). Indeed it did not, since that study was not a study of teacher expectancy effects at all! Rather than manipulate teachers' expectations to determine the effects on pupils' performance, the investigators asked teachers to estimate the future academic performance of black or white boys. The finding of a nonsignificant relationship between children's race and teachers' estimates of future academic success was interpreted as a failure to find an effect of teacher's expectation on pupils' IQ. This study had nothing to do with either IQ or achievement. Teacher expectation was not an independent variable at all but a dependent variable.

Of a total of thirteen studied listed by Jensen (p. 608) as showing no "effects of teacher expectancy on children's IQs," four of the studies (or 31 percent) did not even employ IQ tests as dependent variables and one (or 8 percent) did not even employ teacher expectations as an independent variable. We expect a given degree of error even in science (Rosenthal 1978); but neither the present rate of making errors (e.g., 31 percent) nor the present rate of bias in these errors (e.g., 100 percent) is within acceptable limits.

The mean effect size of the thirteen studies listed by Jensen is 0.00 σ units (Cohen 1977), the mean standard normal deviate (Z) corresponding to the level of statistical significance is .04, and none is significant at the .05 level. Table 1 compares this set of thirteen studies made up of studies of teacher expectancy effects on IQ (N = 9) studies of teacher expectancy effects on achievement (N = 3), and a study of teacher expectancy effects as an outcome variable (N = 1), to several other sets of studies reported by Rosenthal and Rubin (1978) including: (1) all dissertations with special controls for cheating and observer errors; (2) all studies, including dissertations, with special controls for cheating and observer errors; (3) all studies of everyday situations including all studies of expectancy effects in classrooms on IQ, achievement, and other pupil behaviors; and (4) all studies examined in our earlier review. Whatever subset is examined, none is as uniformly negative in its results as are the thirteen studies selected by Jensen. When we examine just that subset of studies previously reported (Rosenthal & Rubin 1978) as most like that of Jensen's thirteen, we find the eight studies of everyday situations conducted as doctoral dissertations with special controls for cheating and observer errors to have a mean effect size of 1.08, a mean Z of 2.35, and six of them, or .75, reach significance at $p < .05$.

Jensen also reports several studies of the expectancy effects of the examiner during the course of psychological testing. The study he singles out as "most powerful and most informative" (Samuel 1977) is one that he feels provides no support for the expectancy hypothesis since "only" 6.4 percent of the variance of that study is attributable to expectancy effects. The error here is in thinking that 6.4 percent of the variance is of little practical consequence. As Rosenthal and Rubin have pointed out elsewhere (1979, 1980), accounting for 6.4 percent of the variance is equivalent to increasing the success rate of a new treatment procedure from 37 percent to 63 percent, a change that can hardly be considered trivial.

Acknowledgment

Preparation of this paper was supported in part by the National Science Foundation.

***by* Robert J. Sternberg**

Department of Psychology, Yale University, New Haven, Conn. 06520

Intelligence and test bias: Art and science

During recent years, an enormous literature has developed on the subject of mental-test bias. Given the importance of this literature to contemporary practices in education, industry, and government, there was a pressing need for someone to undertake the herculean task of assessing and integrating this diverse literature. Arthur Jensen undertook the mission, and deserves credit for his willingness to pursue a task that few people would have wanted to undertake, or, for that matter, could have undertaken. For the most part, Jensen simply compiles what is already in the literature. His book represents neither a radical nor a reactionary reinterpretation of existing data; rather, his interpretations are for the most part consistent with those of the authors who have conducted the studies.

I believe that Jensen has made a serious effort to be fair and balanced in his review. The book will be attacked for all of the wrong reasons—such as that the conclusions are unpopular and that Jensen has proved to be a suitable target in the past—mostly by people who do not bother to read the book. I am concerned that these attacks will divert attention from what I see as the major unresolved issues concerning test bias. I would like to call attention to what I believe are some of the important issues.

First, *bias* can refer to many different things (something of which Jensen himself is obviously aware), at least two of which are particularly relevant here. One meaning of bias is the rather narrow statistical one that comprises the focus of the book: "In terms of predictive validity, a test is defined as biased with respect to two (or more) groups when either the regression of the criterion variable on estimated true scores or the standard error of estimates, or both, are different (i.e., a statistically significant difference) for the two groups" (p. 454). Jenson concludes that there is virtually no evidence that mental tests are biased in this way. In general, I concur with this conclusion, although I am convinced that the tests do discriminate against individuals with test-taking difficulties of one kind or another, and that this discrimination does not manifest itself when data analyses are conducted on groups.

The second meaning of bias is the broad cultural or societal one. Our social system systematically instills in certain people, but not in others, the mental sets and skills that lead to reduced test scores. That such bias exists is compatible with Jensen's own view that "no really clear distinction can be made operationally at the level of tests between *intelligence* and intellectual *achievement*, although intelligence and achievement can be clearly distinguished at the conceptual or theoretical level" (p. 250). At present, at least, we have no way of measuring intelligence except through tests that, at one level or another, are achievement tests. Often, intelligence tests measure achievement for things one should have learned some years earlier, whereas achievement tests measure achievement for things one should have learned in the more distant past. One has only to look at the gross inequities in schooling across this country to realize that children's opportunities for intellectual achievements are not distributed equally. This broad source of bias does not show itself clearly in psychometric studies of test bias because, as Jensen realizes, the bias exists to a large extent in the criterion measures as well. If one's goal is to examine test bias only in the narrow statistical sense, then this broad source of bias will not be of interest. If, however, one's goal is to examine bias in the societal and cultural sense, this broad source of bias will be of interest, and the question will become one of how people

can be trained, on the one hand, and jobs and schooling redesigned, on the other, to instill an equity in our social system that presently isn't there. Jensen cannot be fairly criticized for not dealing satisfactorily with this question, because it is outside the scope of his book. But the reader should recognize Jensen's narrower focus, realizing that there is more to bias than the book might lead one to believe.

Second, it is important to keep in mind the distinction between intelligence as a broadly conceived entity and intelligence as the much more narrowly conceived entity measured by IQ tests. Jensen's chapter asking whether "IQ tests really measure intelligence" (chapter 6) provides a very good review of the concept of intelligence as a broadly conceived entity, one that takes into account adaptation in its many forms to the natural and human environment. But Jensen ends up more satisfied than I am that the mental tests we use give us a reasonable measure of this broadly based intelligence. My own conclusion is that the tests give us a reasonable measure of a fairly narrow subset of the abilities that constitute intelligence in all its manifestations. Although we may not yet be able to measure reliably and validly the other abilities that constitute intelligence, and although we may not even be sure at this time what they are, I don't think we should be quite so satisfied with what we have. I agree with Jensen that we do not yet have a satisfactory substitute for mental tests, but I think there is a clear need to do the research that will enable us to devise tests that more faithfully represent a sampling of behaviors that matter in real-world settings. What investigators of test bias have pursued so far, then, is a narrow conception of test bias as it pertains to a narrow conception of intelligence.

Third, it seems potentially misleading to speak at this point about "culture-reduced tests" (p. 374).[1] First, even the content of the so-called culture-reduced tests favors the kinds of abstract geometric forms that are popular and commonly used in our culture but not necessarily in others. Second and more importantly, the broader kind of test bias I considered above is built into the very structure of the test-taking situation. One's motivation to please the examiner, to solve problems that are in many ways silly and irrelevant to real-world concerns, to solve such problems by oneself, and, in general, to deal with the test and test situation in the terms that it is presented, are cultural norms that are not shared by all elements in our society or other societies. Cole et al. (1971) suggest that "cultural differences reside more in the situations to which particular cognitive processes are applied than in the existence of a process in one group and its absence in another" (p. 233). If this is the case, and cross-cultural research seems to indicate that it is (see also Goodnow 1976), then bias of the more general kind is built into the testing situation as it is now constituted. This bias reflects the values of one particular cultural group, namely, our own.

Fourth, I believe that the test-bias literature that Jensen reviews looks in the wrong places for bias, even bias of the narrower statistical kind. Researchers have tended to concentrate on what E. Gordon (in press) and his associates refer to as "status variables," variables such as sex, ethnicity, race, and socio-economic level, rather than on "functional variables," which include cognitive styles, motivations, problem-solving skills, and personal preferences in working situations, among other things. If there are score differences between status groups, they almost certainly derive from functional variables, and we will not understand the differences until we understand the functional variables that mediate these differences. Boykin's research (see, e.g., Boykin, in press) represents a promising start in this direction.

Finally, Jensen's generally balanced presentation is marred by what I believe are occasional gratuitous statements that are potentially contentious and not well supported by the available evidence. These statements could easily have been omitted without reducing the contribution of the work as a study of test bias. In fact, their omission probably would have increased the impact of the book, because they, rather than the bulk of the work, are likely to be the main objects of criticism. An example of such a statement is that "the white-black difference is mainly a difference in *g* rather than in groups factors that are specific to certain types of items" (p. 585). This statement puts one in the position of choosing between two alternative sources of a difference in mental-test scores (*g* versus group factors), neither of which seems in fact to be the source of the difference. Whereas there is strong evidence in support of social class bases for group differences in test scores (which are still in need of explanation in functional terms), the evidence for racial bases for group differences is very weak indeed (see Scarr & Carter-Saltzman, in press).

To conclude, *Bias in Mental Testing* represents a commendable review of a large and complex literature, albeit a review that is marred in certain respects. I believe that on the whole Jensen's book represents good science, but science conceived of within a narrow frame of reference. And the field that is the object of the book's inquiry is one in which it is essential to have the broadest possible frame of reference. Jensen's narrow conception of test bias represents the state of the field, and thus Jensen's review is faithful to the field as the workers within it conceive of it. But the field isn't quite doing its job. What is missing from the book is an assessment of how the research program of the field as a whole is lacking, and of how this program of research could be improved. It could be improved, I believe, by taking a broader view of the notions of intelligence and of test bias. Alternative vehicles for assessing intelligence need to be explored, and a more humane view ought to be taken of what psychometricians could and should be accomplishing – the maximal utilization of the full range of human potentials.

Recent attempts to exploit Vygotsky's (1978) concept of a "zone of potential development" (e.g., Brown & French 1979; Feuerstein 1979a, 1979b; Laboratory of Comparative Human Cognition, in press) represent a promising lead in this direction, although it is much too early to assess whether these attempts will fulfill their promise. In particular, the concept of a "zone of potential development" itself still stands in need of construct validation. If we are to understand test bias fully, we must ask how well tests are doing in terms of the broader goals of society as well as in terms of the narrower goals of test publishers and some psychometricians. The narrow psychometric view of equity is a not unimportant one, but the broader societal view is the more important of the two.

Acknowledgments

Preparation of this report was supported by Contract N0001478C0025 from the Office of Naval Research. My views on test bias have been shaped in part by discussions held in the developmental psychology postdoctoral seminar held at Yale University, and in particular by the contributions of Tom Berndt, Ed Gordon, Bill Kessen, and Sandra Scarr, who of course are not responsible for any of the statements I have made here.

Note

1. Jensen properly rejects the terms "culture-free" and "culture-fair" as misleading.

Editorial Note

A *BBS* treatment of "Sketch of a Componential Subtheory of Human Intelligence," by R. J. Sternberg, appears in the next issue, *BBS* 3(4) 1980.

***by* Leona E. Tyler**

Professor Emeritus, University of Oregon, Eugene, Ore. 97403

Tests are not to blame

Bias in Mental Testing is a most impressive book. Jensen has extended his search for relevant evidence backward to the early years of the century and forward to the time the book went to press. Data from studies psychologists have forgotten for years are brought into alignment with data from recent studies utilizing the most sophisticated statistical and computer technology. Government reports, Army and Navy publications, and other sources generally ignored in research surveys are mined for the information they contain. The discussion is a model of clear exposition. Concepts often confused, such as "bias" and "unfairness" are clearly differentiated. Complex statistical techniques are explained in simple language. Examples are plentiful and cogent. Graphs are meaningful. You may not like the author's conclusions, but you cannot lightly dismiss them.

Basically, what Jensen has done is to reinstate a theory that was fundamental to differential psychology during its early years, Spearman's *g* theory, showing that recent as well as older evidence supports

the assumption that all mental tests measure *g* along with whatever else they measure. Spearman defined this ability as the *eduction of relations and correlates*, and Jensen has marshalled the evidence that it is this characteristic that accounts for the differences between races and social classes.

If Spearman's principle of "indifference of the indicator" holds true, test bias is not an important factor. Whenever several tests are given, the specific factors cancel out, so that *g* is what is measured. Jensen's case that group differences represent *g* rather than test bias rests on both internal and external evidence. Factor analyses support the conclusion that the pattern of relationships between scores is similar in many specific groups tested, and thus that what is being measured is similar in the various groups. Regression equations for predicting various criteria can be compared in three ways, slope, intercept, and standard error of estimate. The consistent difference Jensen finds between blacks and whites is in *intercept*. Its direction is such as to favor blacks over whites in selection situations. Criterion scores such as college GPA or occupational ratings tend to be overpredicted for blacks, underpredicted for whites.

If we accept this evidence for a conclusion which is the opposite of what the test bias hypothesis would lead us to expect, and it appears incontrovertible, we are left with the fact that some groups in the population average higher than others in intelligence. Other ways of avoiding this conclusion are also shown to be untenable, such as the idea that examiner or teacher expectancies [see Rosenthal & Rubin: "Interpersonal Expectancy Effects" *BBS* 1(3) 1978] lead to higher scores for some children than for others, or that people score higher when tested by a member of their own race. And there is abundant evidence that test items that look inappropriate are not necessarily invalid.

I accept Jensen's conclusion that mental tests in common use are not biased against disadvantaged groups. It follows that invectives against testing or legislation prohibiting it will not eliminate the disadvantage. In making this point clear for all concerned, the book performs a useful service. It is its treatment of the implications of this conclusion that I found most unsatisfying. What must be recognized is that the intellectual climate of the 1980s is vastly different from that of the 1920s when the *g* theory was proposed. Then class lines were sharply drawn in Great Britain and only slightly less sharply in the United States; good Negroes "knew their place," and only the brightest or most privileged students expected to go to college. Whether we like it or not, that world is gone, and we are committed to an egalitarian society that will provide not only equal opportunity but equally rewarding lives for all of its members. Individuals or groups are no longer willing to be labeled "inferior."

Some research directions Jensen fails to emphasize seem to offer more promise for progress toward this ideal than continued attempts to document the existence of *g*. One is the direct attempt to accelerate the development of intelligence in young children. The large and significant body of work by J. McV. Hunt (Hunt 1979) shows conclusively that the rate of mental growth can be speeded up in early childhood. It finds no place in Jensen's pages, although it would tie in well with his conclusion that the black-white difference is essentially a difference in mental maturity. Another unemphasized research direction is the exploration of the *non-g* aspects of human functioning. Jensen himself has done some of this work, and had he labeled his two kinds of ability something other than Level I and Level II, which imply that one is higher than the other, the distinction between intelligence and the ability to learn a wide variety of useful information and skills might have been more widely accepted. Similarly, the significance of Guilford's (1967) pluralistic approach to the problem of human ability is minimized.

In his last chapter Jensen does discuss some of the implications of his findings for schools and for society. But changes far more drastic than any he proposes will be needed if they are to have an impact on the problem of inequality. What one hopes to find in an author with Jensen's awesome grasp of research-based knowledge is more imaginative speculation about what might be done. Perhaps this book is not the place for such an exploration of possibilities, but it should be undertaken somewhere, by someone.

by **Steven G. Vandenberg**

Institute for Behavioral Genetics, University of Colorado, Boulder, Colo. 80309

An existence proof for intelligence?

This is a magisterial book which concludes that there is no bias against racial or socioeconomic groups. Jensen begins by reviewing popular and occasionally professionally held prejudices and misconceptions about intelligence per se – including the notion that one cannot measure something that cannot be defined, or that does not exist as a real entity or force.

In 740 pages with 28 pages of references, Jensen builds a strong case for the existence of a human attribute which can be assessed by a variety of mental tests, which, most likely, is normally distributed in the population, and for which the applicability of an interval scale can be defended. His arguments are presented in great detail because critics of the use of intelligence tests frequently sidestep the issue of bias against minorities by questioning the very basis of testing when they claim that intelligence has not been – and cannot be – defined. It is true that intelligence has far too often been operationally defined, "intelligence is whatever intelligence tests measure," but it can hardly be said that Binet or Spearman did not present theoretical definitions. Nevertheless, the existence of "intelligence" is often simply taken for granted. When critics say that there is no adequate theory, what they really mean is that there is not a *complete* account of the development of intelligence, its underlying structure, its everyday manifestations, and perhaps its physiological underpinnings; all there is instead is a statistical methodology ("psychometrics") which, according to the critics, is without a real scientific basis. Aside from the fact that one could say the same about gravity, genes, and most scientific concepts, many of these assertions are incorrect. It just happens that most texts about tests and testing, or even those about test theory or measurement, do not address these issues. They are to be found instead in texts on child development, especially in discussions of Piaget's work, and in physiological psychology texts or in books about cognitive psychology. Jensen has performed a real service by pulling together a number of approaches used to establish a scientific basis for the concept of intelligence – almost an existence proof. [See also Sternberg: "Sketch of a Componential Subtheory of Human Intelligence" *BBS* 3(4) 1980.]

One might have wished more of a developmental point of view and particularly more use of Piaget's ideas in this context, especially since Jensen mentions Piaget's work with admiration later in the book in the context of alternative methods of assessing intelligence and as almost circumstantial evidence of the existence of intelligence, rather than as the core of the theory. In fact, it may be telling evidence of Jensen's views that his admiration for Piaget's work seems to some extent to derive from the remarkable psychometric properties of some scales constructed, not by Piaget, but by others.

Throughout the book the theory of a unitary, general intelligence factor, *g* in Spearman's tradition, is defended, with special abilities playing very much a subordinate role. Jensen does not fully address the objection that this *g* varies as a function of the various tasks in a given test and the various tests in a battery. This surprised me, because such variability can account in part for the lack of constancy of an individual's test scores, especially during childhood. As a consequence of this favoring of *g*, Jensen makes the rather startling statement that it is wrong to include the first principal component when rotating factors orthogonally, as for instance by Varimax. According to Jensen, there are only two correct procedures when analyzing the correlations among ability tests: (a) one uses an oblique solution followed by a second-order factor analysis, or (b) more simply, one leaves the *first* principal component unrotated (i.e., just as it was found) and interprets it as the *g* factor, while rotating as many of the remaining components as one would otherwise (based on the size of the eigenvalues) *plus one*. These rotated factors are then the "lower order" specific ability factors *minus* their contribution to the general factor. While one can understand Jensen's preference for this procedure, it seems rather harsh to say that a position taken by Thurstone, Guilford, and many others is wrong. It is true that it makes no sense to

rotate to an orthogonal solution and then to use this as "proof" that no general factor exists. But authors should be permitted to use an orthogonal solution if they make it clear that this is their preference and perhaps state why they prefer this solution.

After all, these are alternate *models*, each of which accounts for the data equally well. The question is, which model is more useful for the particular purpose one has in mind. A general factor may be best for predicting success in school, while differential diagnosis will be better served by a model with orthogonal multiple factors. The work by Sperry and associates on split brain patients suggests that there are at least two relatively independent types of abilities. I have elsewhere reviewed other evidence for the usefulness of specific abilities versus general intelligence (Vandenberg 1968, 1973) and concluded that the evidence is inconclusive, but this is partly due to the scarcity of well-designed studies in which specific ability measures were used.

The question of how to derive factors is especially crucial for behavior genetics. In searching for separate genetic mechanisms for specific abilities, one does not want to use scores that are "contaminated" with *g*, because that might lead to more similar estimates of genetic variance for each specific ability and make it harder to establish genetic specificity. On the other hand, each step which removes variance moves one further away from the actual individual observation. For that reason it is better to proceed directly to the elimination of genetic and nongenetic correlations from the raw scores (Vandenberg 1965; Bock & Vandenberg 1968; DeFries, Kuse, & Vandenberg 1979; Loehlin & Vandenberg 1976).

I will not review the question of bias, even though that is the central concern of the book. I trust that others will comment on that issue. I, for one, was already convinced that the only bias of some ability tests is against poor education to the degree that those tests call for prior exposure to experiences that are either deliberately or incidentally produced by education. The evidence accumulated by Jensen about highly similar rank orders of item difficulties and highly congruent factor structures, combined with the fact that initial mean differences are small, suggests that more detailed longitudinal studies of the relative decline of poor students are needed. Even so, this cannot properly be called bias, because the tests are doing what they are designed to do: measure present performance as a predictor of later performance. Such findings call for changes in education, not in testing.

I have a few minor comments, as follows:

1. More recent studies arrive at estimates of genetic variance lower than 80 percent, especially when they have used corrections for the effects of age, which tended to inflate previous resemblances. (p. 58)

2. The assumption of a normal distribution of intelligence is valid only within the range of testable human intelligence. There are no adequate methods for assessing the intelligence of severely retarded individuals, such as those who would receive virtually zero raw scores on the WISC test. And, of course, there are no methods at present which permit comparing human and animal intelligence. At best, one can say that an IQ of 100 is quite high on an absolute scale with a true zero. (p. 75)

3. That there exists no competing theory about the distribution of intelligence is true as far as a complete statement goes, but it may be possible to develop one from absolute performance measures such as size of vocabulary, ability to mentally rotate figures of increasing difficulty, and so on. Most of this type of research still needs to be done and, unless a pressing need for it develops, it will not be done. However, the absence of data cannot be taken as an indication that it could not. In the meantime, a number of possibilities have been suggested, even for cross-species comparisons, which might produce highly skewed distributions or stepfunctions. (p. 87)

4. Genetic correlations have been separated from nongenetic correlations. For a review, see DeFries, Kuse, and Vandenberg (1979). Adoption studies permit even better separation. (p. 58)

5. In the large sample used in the Hawaii Family Study, a small but significant sex difference of one point in favor of males was observed for the Progressive Matrices (Wilson & Vandenberg 1978). (p. 647)

6. Rulon's test was not released by the Army for research by outsiders; at least when I enquired some years ago, permission was not given. I agree that it would be worthwhile to try again. (p. 654)

In conclusion, this is a scholarly work of tremendous importance. It is to be hoped that the considerable technical detail will not deter a large percentage of the audience which ought to be reached from reading this book. Now that the use of testing has become a matter for the courts, it is especially important that judges become familiar with the evidence on which the conclusion that there exists no bias in mental tests is based.

***by* P. E. Vernon**

Department of Educational Psychology, University of Calgary, Calgary, Alberta, Canada T2N 1N4

Antitest views are refuted

There are few psychologists nowadays who would be capable of writing a book of about half-a-million words, while heavily involved in university teaching, in carrying on research and publishing several articles each year, and in coping with the endless stream of brickbats flung by his critics. On every score this book demands the attention both of Jensen's supporters and his opponents. Indeed, it has already received a good deal of publicity from the press and media, which is somewhat unfortunate, since the commentators appear to assume that this is Jensen's reply to eleven years of criticism, and that its main aim is to establish the existence of racial (genetic) differences in intelligence. But in fact Jensen does not discuss negro inferiority in intelligence, or the reasons for the lower scores of blacks than whites. The nearest he comes to this issue is the following sentence from a footnote (p. 58): "The idea of a genetic component in the racial (i.e., black-white) IQ differences is the most disputed and at present is generally regarded by geneticists as a scientifically legitimate but unproved hypothesis."

In other words, Jensen is a good deal less extreme in his views about race differences than he appeared to be in the famous 1969 article. Elsewhere he goes to considerable lengths to point out that test scores or IQ differences tell us nothing about causes. What such scores do show is a matter for empirical research, whereas explanation is a far more complicated matter, which will require the combined efforts of biological and behavioral scientists to unravel. He also still holds that there is strong genetic influence in individual differences in ability, but admits that results vary, and that the mean heritability coefficient is probably about 0.7 rather than the 0.8 claimed earlier.

Jensen's main purpose is to investigate the common allegation that intelligence tests are made up for, and standardized on, whites (mainly middle-class), and are therefore unfair to and biased against members of minority groups, or lower working-class children and adults. The first couple of chapters document this allegation very fully and quite objectively. In later chapters the nature of bias, of cultural loading, and validity are very thoroughly analyzed, and an enormous array of evidence is summarized which contradicts the accusation. Cognitive tests are shown to measure the same abilities among all English-speaking cultural subgroups brought up in the U.S. Under some circumstances, as in tests used to select students for college or adults for jobs, there is often a small amount of bias, but this favors, rather than disfavors, black applicants. That is, the tests tend to overestimate their suitability for higher education or skilled jobs. Doubtless a lot of readers of the book will still not be convinced. But if they do follow the arguments and the results of hundreds of investigations, they should realise that their antitest views are untenable.

Jensen admits that the correct interpretation of statistical evidence is highly technical, and he therefore includes six chapters (4–9) of what is virtually a textbook of psychometrics. Probably this is the most advanced book on test theory and statistics since Harold Gulliksen's thirty-year old text (Gulliksen 1950). I am a little afraid that this will put off most readers who do not already have a considerable background in statistics. True, this plays an important part in the proper interpretation of the evidence presented in later chapters, but I doubt if it is essential to comprehension. Fortunately a shortened and simplified

version of the book is now ready for publication, and this should certainly attract and influence a much wider audience.

by **Atam Vetta**

Oxford Polytechnic, Oxford OX3 OBP, England

Correlation, regression and biased science

Crow (1969), commenting on Jensen (1969), made a public affirmation of his admiration for Jensen's knowledge of genetics, and the latter (Jensen 1969a) made full use of this accolade from one of the leading quantitative geneticists. When I read Jensen, I felt it necessary to say that "the very small sections of his work that concern genetical concepts show some confusion and are, in places, totally inaccurate. It is, therefore, important that researchers in fields of genetics and IQ be aware of the deficiencies of his excursions into genetics" (Vetta 1977). On reading Jensen's assertion in the preface of his book that "Anyone who would claim that all such tests are, therefore, culturally biased will henceforth have this book to contend with," I thought he might have something credible to offer. I was fortified in my hope with the knowledge that he has worked in the field of mental testing for quite some time. It is with sorrow and regret that I report that the strictures quoted above apply with greater rigour to his present work, *Bias in Mental Testing*.

A reader has to wait till chapter 9 to find the definition of bias. The first few chapters are a partisan exposition of some events and views. Jensen is as entitled to his prejudices as I am to mine. I would not, however, describe such an exposition as "scholarship," as Dr. Stanley does on the jacket. The book is replete with inaccurate statements and there appears to be confusion concerning the meaning of some concepts and formulae. This review contains a sample of these. Before discussing them I would like to say that I deprecate the practice of combining results from two separate studies which use different tests, in an analysis of variance table as given on p. 43. [Cf. Rosenthal & Rubin, *BBS* 1(3) 1978].

Jensen asks (p. 75): "So how can we ever make sure that the test scores represent an interval scale?" His answer is that is we *assume* that the distribution of IQ scores is normal and "if we can construct an actual test that in fact yields a score distribution like the one we have assumed, we can be absolutely certain that the scores are on an equal-interval scale . . . . *Ipso facto*, any test of intelligence that yields a normal distribution of scores must be an interval scale." I regard this last statement as rather naive. Actually, Jensen's earlier position (1969) was slightly more defensible. Then he cited the pattern of kinship correlations as an objective test and prompted the retort from Hunt (1969), "Am I emitting a mere flippancy if I respond that apparently, for Jensen, going twice around the circular argument removes its circularity?"

Jensen cites the regression of one sib's IQ score on that of the other as crucial evidence in favour of the polygenic theory and, hence, of the interval scale. He appears to have discarded the regression of progeny mean on mid-parent as crucial evidence, in favour of sibling regression. His work shows confusion concerning the significance of regression. Thoday (1973) offered the correct interpretation of the regression among black and white siblings, and, even though Jensen does not acknowledge his debt to Thoday, he has now noted that a population must regress on its own mean. I appear to have been less successful in convincing him that regression on the mean, in the absence of other evidence, provides no proof of the polygenic hypothesis or the interval scale (Vetta 1975). Regression is the statistical consequence of imperfect correlation between two variables. It is incorrect to cite any type of regression as a proof for polygenic hypothesis.

Jensen's statement (p. 184) that "Genetical models . . . fit the various . . . correlations for IQ remarkably well (Jensen [1972] . . . Jinks & Fulker 1970)" can be challenged. Jinks & Fuller use Fisher's (1918) model of assortative mating on Burt's data; the latter are now utterly discredited. Moreover, the observed correlations for IQ for a given kinship vary so much that they cannot be regarded as having come from the same population. The parent-child correlation, for example, varies from about 0.2 to about 0.8. Jensen (1969) regards it as providing "compelling" evidence for the genetic hypothesis. Vetta (1980) says that "The 'compelling' evidence dissolves at the first sight of a statistical test of significance."

Jensen offers (p. 428) a test for whether IQ scores follow an interval scale and says that "The hypothesis of an interval scale can be rigorously tested by determining if there is a significant correlation between the sibling means and [their absolute] differences." If the correlation between the two is zero, then IQ follows an interval scale. Should anyone be tempted to quote this test, I must point out that I consider it to be without validity. It is a rehash of Jinks & Fulker's (1970) test for Genetic/Environmental (G × E) interaction. Even though Dr. Herrnstein [then editor of *Psychological Bulletin*] refused to publish my correction to Jinks & Fulker (in spite of the fact that Dr. Jinks accepts my criticisms), it could not be unknown to Jensen. I am, therefore, surprised that he puts so much credence in this test. Perhaps he did not understand the full implications of my correction. The fact is that if the distribution of a sibling on IQ is normal and that of both siblings is bivariate normal, the theoretical value of the correlation between sibling means and their absolute difference is zero. Thus, using Jensen's test we obtain zero correlation because we have made the distribution of IQ normal. No other interpretation is possible. (A copy of my paper on Jinks & Fulker may be obtained from me on request).

Jensen discusses (pp. 243–245) broad heritability and asserts that most behaviour geneticists agree that heritability of IQ is substantial. If they do, it is regretable, and indicates a deplorable lack of knowledge on their part. They should be advised to read Kempthorne (1978) so that they may understand that without breeding experiments we cannot make claims concerning genetic determination. Observational studies of the type conducted for IQ can, at best, enable us to claim an *association* between a trait and "hereditary factors." These hereditary factors must of necessity include those environmental factors which increase resemblance between relatives. It would also be useful to contemplate the reasons why Fisher (1918) assumed random environment. The fact is that without this assumption his definitions of additive and dominance deviations will not hold, and if they can still be defined they have little genetic significance.

The errors of measurement associated with IQ cause me some concern and I am surprised that they evoke no questioning from Jensen. Fisher (1918) assumed no errors of measurement. It was Brown's (1930) study (p. 283) which first made me wonder about the applicability of Fisher's theory to IQ. Assume that mean IQ = 100 and its variance = 225. Jensen's table 7.11 (p. 284) shows that error variance (i.e., the variance of the score difference on two occasions) is $(12.99)^2 = 166.4$ (yet he takes the error variance at 5 percent, i.e., 11.25, on p. 43). 166.4 is 74 percent of 225. Given this proportion of error variance, what credence can I have in the statement that 75 percent of the phenotypic variance of IQ is due to genetic causes?

Jensen provides (p. 279) a table from Hirsch (1930). This and other studies of that period are rather interesting. They show that other ethnic groups such as Irish Americans, Italian Americans, Portugese Americans, and so on, had an average IQ not far above and, indeed, in some cases below the average of black Americans. We do not hear much about these ethnic groups. This may be because (1) their average IQ has increased and is now equal to the average of the other white Americans, or (2) it is politically too dangerous to talk about their low IQ. Jensen will perhaps tell us which of the two reasons is the correct one.

The concept of cultural bias in IQ is well known and understood by everyone. Jensen redefines it. I do not accept his definition and believe that he should have chosen a different name for his own concept, which I feel has no validity. Nothing but confusion results when you take a widely used concept and give it a different meaning. Jensen is not really discussing cultural bias but his concept of "predictive test bias." A short paper in a research journal explaining this concept should have been sufficient. He would regard a test as having predictive bias if its regression coefficients, intercepts, or standard errors for the two groups differ (p. 381). He then hypothesizes different mental growth

rates for the two ethnic groups (the idea of polygenes causing different mental growth rates in ethnic groups of a nonisolated population may prove to be too much even for those geneticists who admire Jensen's understanding of genetics). He then asserts that on this basis we should find $\bar{X}_b/\bar{X}_w = \sigma_b/\sigma_w =$ a constant. Without using a test of significance he says that the difference between the two ratios which equals 0.03 "is a nonsignificant difference" (p.425). So the hypothesis of different mental growth is acceptable.

I do not accept Jensen's reasoning but would like to point out that a proper test of significance would show that the two ratios differ. The equation involving the two ratios can be written in a slightly different form, namely, $100\ \sigma_b/\bar{X}_b = 100\ \sigma_w/\bar{X}_w$. This, of course, means that the two coefficients of variation are equal. To avoid the complexity of finding the standard error of the difference between the two coefficients of variation, we may use the simpler technique of confidence limits. Black sample size was 1800. The black and white coefficients of variation, V_b and V_w are 15.37 and 16.11 respectively. The standard error of V_b is 0.256 and its 95 percent confidence limits are 14.87 to 15.87. V_w lies outside these limits. The two ratios differ significantly. Jensen's assertions consequently cannot be accepted.

I am impressed by Jensen's faith in correlation and regression formulae. Perhaps, for him, the manipulation of these formulae is a substitute for further analysis. I do not believe that analysis of a correlation matrix, by whatever method, can deepen our understanding. I find it odd that some psychologists believe that scientific thought ended with Spearman's *g*.

Bias in Mental Testing is intended to be a defence of mental tests and in spite of many emotional appeals, it fails to achieve its purpose. Jensen asks: what is the alternative? I find the question rather odd, because here in England we rarely use IQ or other mental tests. School-leaving certificates (Ordinary and Advanced level) are regarded as sufficient for almost all purposes. To many of us, excessive testing is an American idiosyncrasy. It is our hope that American society will grow out of it. Meanwhile dare I ask: is it really beyond the American genius to devise a system like ours?

***by* F. Vogel**

Institute for Anthropology and Human Genetics, University of Heidelberg, 6900 Heidelberg 1, West Germany

Genetic influences on IQ

A. Jensen's work is controversial mainly because he has interpreted the undisputed group differences in the outcome of IQ tests, primarily those between American blacks and whites, as indicating corresponding genetic differences. It is my contention that the primary cause of this controversy is not ideological prejudice on the part of either Jensen or his critics but the inherent weakness and low explanatory power of the paradigm of quantitative - biometrical genetics on which his interpretation is based.

As Jensen correctly states (p. xi): "Test scores . . . are measurements of *phenotypes*, not *genotypes*." The genotype, however, determines the phenotype in interaction with the environment in the course of individual development. This statement, trivial as it sounds, leads the geneticist to investigate specific kinds of interaction, and especially the role of genes in this process. Jensen writes (p. 183). "The genotype is itself a theoretical construct." This is true as long as we treat the genotype as a whole in our analyses, and attempt to measure its contribution to the interindividual variability of a parameter, for example the IQ. This approach is no longer necessary any more, however.

The development of concepts and research techniques in neurobiology, biochemistry, and molecular biology has resulted in much more penetrating analyses of the biological mechanisms of gene action, for example, in medical genetics. Interactions of specific genotypes with specific environmental influences have been discovered (Vogel & Motulsky 1979; *Human Genetics, Problems and Approaches,* Springer Verlag). Admittedly, application of these principles to the problem of genetic differences in intelligence meets with technical difficulties, and has been attempted so far only occasionally. There is little doubt, however, that progress in this field is forthcoming. The "black box" of the genotype will gradually be broken up.

As Jensen states (p. 740): "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." It might be a good idea to postpone further discussion of the relative proportion of genetic and nongenetic factors in the observed variability of intelligence and performance, and on a possible genetic component in the observed group differences to a time when the genetic mechanisms involved will be better known. In my opinion, even Jensen's statement (p. 244) that "it would be difficult indeed to make a case for the hypothesis that the heritability of IQ is less than .50" is too optimistic in view of the inherent shortcomings of the quantitative-genetic method – shortcomings that are mentioned in part by Jensen himself. On the other hand, it would be surprising if there were no genetic variability in genes influencing human behavior and performance in the normal range; such a contention contradicts all experiences in other fields of genetics.

These problems, however, are not the main topic of Jensen's book. In recent years, criticism of Jensen and his conclusions has developed into criticism of intelligence and personality testing in general. Tests have been accused, for example, of examining only the presence of white middle-class concepts, and, in general, of being biased against minorities, especially American blacks. Being a human geneticist, not a psychometricist, and living in a society in which the problem of ethnic minorities is (still) relatively minor, I feel unable to assess competently all the details of Jensen's extended and penetrating technical discussions. However, I share his opinion that good tests indeed measure abilities that are important for predicting success in school, college, and professional careers. Moreover, some recent results in behavior genetics have shown that, in principle, these tests are sensitive enough to indicate even relatively small deviations caused by known biological mechanisms; examples are the specific defect of spatial orientation in Turner's syndrome, a slight verbal weakness in heterozygotes of the phenylketonuria gene (Thalhammer et al. 1977; *Human Genetics* 38:285) or a weakness in the performance part of the Wechsler test in heterozygotes of the ornithine transcarbamylase deficiency (Batshaw et al. 1980; *New England Journal of Medicine* 302: 482). These results hint that there may be much more widespread causes for genetic influences on the IQ in the "normal" range.

In my country, IQ and personality testing is not nearly as widespread as in the USA. Selection for school curricula, university admission, and job assignment is normally carried out much more informally. This might occasionally help to meet individual demands better; however, I have very serious doubts whether our (largly systemless) system is, indeed, more efficient and, above all, fairer than the ubiquitous use of tests which have been tempered in the fires of criticism.

***by* Douglas Wahlsten**

Department of Psychology, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1

Race, the heritability of IQ, and the intellectual scale of nature

I would like to address several questions pertaining to this book.

1. What is a race? The book deals primarily with the notion that different average IQ scores for different "races" in the United States reflect cultural or test bias. However, the terms "race" and "racial" difference are used improperly throughout the book. The reader is exposed to data on the "white," "black," "Mexican-American" and "Asian" racial groups.

As a scientific term, as opposed to a colloquial expression, "race" is precisely defined to be a reproductively isolated, stable, and genetically distinct subpopulation of a species. There is no "white" race anywhere on the globe. One could say there is a Caucasian racial stock, but this grouping includes the people of Iran, Afghanistan, and northern India who have various shades of brown skin. Neither is there a "black" race. Africa contains a number of different ethnic groups, some of which have Mongoloid ancestors. People with very dark skin can trace their ancestries to various parts of either Africa, Melanesia, India, or Australia. The idea of a "Mexican-American" racial group is

very difficult to accept. The ancestors of the present-day population of Mexico include the Caucasians of Spain and the native people of Mongoloid extraction. There has been substantial reproductive intercourse between men and women of European, African, and Asian ancestries in the United States, which further blurs any ancient racial boundaries that may have existed.

The author makes no mention of how "black" and "white" were defined in the various studies he cites or how people with parents from different ethnic groups were assigned to one group or the other. It seems to me that if one is seriously interested in the relation between skin color and performance on various tests, which is an enterprise of dubious merit in the first place, then at the very least one should use a light meter to actually measure reflectance of each person's skin.

2. Is IQ normally distributed? The author argues persistently that mental ability is polygenically determined and therefore normally distributed in the population whereas achievement is skewed because it is a product of "abilities, disposition and training." He claims that for the American population in particular "the normal distribution of intelligence is probably the most unrivaled theory in all of psychology" and uses this conclusion to bolster his contention about genetic determination of IQ.

However, the author does not even mention the excellent work of Dorfman (1978), who has demonstrated beyond reasonable doubt that the IQ distribution is not normally distributed; that IQ deviates more from normality than does height; and that a virtually perfect normal distribution of IQ scores reported by Cyril Burt was fabricated. Jensen further expresses his own bias when he calls an obviously nonnormal set of IQ scores reported by Wechsler "a slightly negatively skewed normal curve" and refers to other instances of skew as "anomalies." If a distribution has statistically significant skew or kurtosis, then it is *not* normal.

The author maintains that IQ for a large sample of American "whites" is "near to normal as can be," whereas for "blacks" the distribution has a lower mean and a slight positive skew. Here Professor Jensen gets into real trouble. Approximately 10 percent of the American population is of African ancestry, whereas about 88 percent is of European ancestry. Pool the "near to normal as can be" IQ distribution for 88 percent of the people having a mean of 100 with the positively skewed distribution of another 10 percent with a lower mean, and you *cannot possibly obtain a normal distribution of IQ for the entire population* in the United States.

Given the diversity of factors which can cause data to deviate from normality, I don't think the shape of a frequency distribution can tell us much about the importance of training for performance on a test of mental ability. Direct measures of experience are necessary, but no such measures are presented in this book.

3. What is the "heritability" of IQ? The assertion by Jensen (1969) that "the best single overall estimate of the heritability of measured intelligence" is .81 has been strongly criticized in articles by Bodmer and Cavalli-Sforza (1970), Hirsch (1970), Lewontin (1970), and many others, as well as in the book by Kamin (1974). The author does not mention or cite any of his critics, and as far as I can determine from his latest book he has not retracted any of his previous views. Instead he states baldly: "Most geneticists who have surveyed the evidence are agreed that some substantial part, probably as much as 80 percent or more of the IQ variance within families (i.e., between siblings) is genetic," and he refers to "the polygenic theory of intelligence, which is now generally accepted by geneticists." I would like to know where this opinion poll of geneticists has been published and who comprises this great silent majority to which Jensen alludes.

The author claims published estimates of heritability of IQ range from about .50 to .90, but he does not mention recent studies reporting data for twins reared in the same homes – where "heritability" is in any case liable to be overestimated because of the shared environment – that suggest values substantially lower than .5 (e.g., Scarr-Salapatek 1971; Wilson 1978).

The author does not even mention the exposure of the fraud committed by Cyril Burt, and instead cites Burt abundantly and relies heavily on Burt's "data" on monozygotic twins reared apart to claim that "the environmental influences on IQ are normally distributed." He also writes: "In any particular study, one can always find methodological reasons for some doubt. The covergence of evidence from many studies using different methods, however, leaves little if any doubt concerning the relatively high heritability of IQ." For one thing, the evidence does not "converge" on a particular number. Furthermore, flawed studies will converge on a *biased* answer if, as Kamin (1974) has shown convincingly, the methodological flaws consistently influence the results in favor of higher correlations among biological relatives.

In my opinion the traditional concept of heritability should be discarded altogether because of evidence that hereditary and environmental factors interact throughout the process of development of the brain and behavior and render simple additive models such as $P = G + E$ invalid (Wahlsten 1979; see also Lewontin 1974). What Jensen has written on genetics in his latest book reveals an extremely narrow and outmoded perspective on the biological sciences.

4. Are humans at the pinnacle of an intellectual scale of nature? The author cites research on learning in "animals at different levels of the phyletic scale – that is, earthworms, crabs, fishes, turtles," and so on, and refers to "the turtle, which is phylogenetically higher than the fish" in order to muster evidence for the "biological reality" of intelligence or *g*.

It is wrong to speak in typological language about "the turtle" or "the fish" when in fact only a very few species sampled selectively from among all living species of turtles and fishes have ever been trained on a learning task, let alone a wide range of tasks which would be necessary to test for the existence of *g* in nonhumans.

Taxonomically, "turtle" is one of four orders in the class *Reptilia*. The contemporary fishes are grouped into three classes – the jawless lampreys and hagfish (*Agnatha*), the cartilaginous sharks and rays (*Chondrichthyes*), and the bony fishes (*Osteichthyes*). Although fish appeared in the fossil record earlier than reptiles, living fishes cannot be characterized as "lower" than reptiles. Fish continued to evolve after giving rise to the amphibians, the ancestors of the reptiles, and many contemporary fishes show intricate social behavior, including parental care of the young, which is lacking in many reptilian species.

The Aristotelean notion of a linear hierarchy of species or *scala naturae* has been repudiated by Hodos and Campbell (1969), especially as it pertains to rankings with respect to the learning ability of various species, but apparently Jensen did not take note of this important paper.

I would also like to point out that two independent reviews of the literature on heredity and learning in animals both concluded that for several nonhuman species which have been extensively investigated, there is no evidence of a general "intelligence" factor, although hereditary influences on task-specific performance are well documented (Wahlsten 1978; Fuller & Thompson 1978).

From my reading of those features of Jensen's book about which I have specialized knowledge, I conclude that this work is so lacking in balance, rigor, and erudition I cannot recommend it, either to professionals in the life sciences or to the general reading public.

Author's Response

by **Arthur R. Jensen**
Institute of Human Learning, University of California, Berkeley, Calif. 94720

Correcting the bias against mental testing: A preponderance of peer agreement

Overview

If any single sentence best sums up the main conclusion of *Bias in Mental Testing*, it is probably the one on the final page, which states, "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). Other summarizing sentences are: "Differential validity [of college and job selection tests given to blacks and whites] is a nonexistent phenomenon. . . . 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 most striking feature of this *BBS* multiple review is the preponderance of agreement with the book's main conclusions – conclusions which, it should be noted, contradict the popular prejudice that standardized tests are culturally biased and unfairly discriminatory against all but the white middle-class. Of the total of 27 reviewers, 18 express agreement with the book's main conclusions. The remaining 9 are either noncommittal or address side issues, but not one directly challenges the main conclusions. This is not to imply that there is a paucity of criticism in this multiple review. But criticism directed at auxiliary points should not be misconstrued as support for the now firmly discredited belief that the disproportions in the selection of different subpopulations on the basis of test scores, in special classes in school, in college admissions, and in jobs, are the result of biased tests. It is especially gratifying to me that so many of the commentators who have directly addressed my book's main findings have themselves made notable contributions to the study of test bias and mental testing in general.

All of these reviews were sent to me just as I had finished reading Horace Judson's (1979) fascinating book on the history of molecular biology, a field that has been characterized by unusually rapid scientific progress, and the research leading up to the discovery of the structure of DNA. As a veteran of criticism, on both the receiving and giving ends, in my own field of differential psychology, I was particularly interested in the rather different tenor of criticism in the field of molecular biology. Most of it seemed to be a highly mutual and usually friendly give-and-take exchange among scientists, all of whom agreed in the main on what the problems were, and were working to solve them and to understand more fully the phenomena of mutual interest. Debate and criticism in what has come to be known as the "IQ controversy," in contrast, has been more ideologically charged and is more often patterned on the polemical style of advocacy, rebuttal, and put-down, rather than being in the spirit of cooperative problem-solving through mutual criticism and correction. The argumentation of "the IQ controversy" reminds one more of the post-Darwinian debates between the evolutionists and the creationists – a controversy that is indeed being revived in the present day.

Except for reviews in the popular press, I expect that less of this style of criticism will attend *Bias* than was seen in the professional literature in past debates about IQ, largely because I have deliberately and necessarily treated the subject of test bias separately from the controversial nature-nurture issue. My expectation is largely borne out in the present multiple review, with a few exceptions. But discerning readers will have no trouble recognizing attempted criticisms of points in *Bias* that depend solely on "guilt by association" and what Ingle (1978), in an article I urge everyone to read, has termed the "poison well fallacy." There are some perfect examples of these in the present collection. In *Bias* (p. 359), I cited six studies that found a negative association between IQ and delinquency, including a 1925 study (probably the first) by Burt. Instead of taking issue directly with this point, **Dorfman** brings up Burt's comments about physiognomy and quotes his description of a boy as a "slum-monkey" resembling a "pale-faced chimpanzee," etc. The tactic is clear. **Hirsch et al.** display a quotation from Spearman supposedly linking *g* with eugenics, which is still an emotive word that upsets some people. Need anyone be reminded that a scientific construct, as *g* is, must stand on its own, regardless of the personal philosophy or motivation of its originator? Should *g* by regarded any differently by scientists even if it had been propounded by, say, Charles Manson instead of by Charles Spearman? **Vetta** refers to Table 7.7 in *Bias* (p. 279), from a study by N. Hirsch (not to be confused with J. Hirsch) showing the year-to-year intercorrelations of IQs for a group of children in grades 1 to 6. Rather than taking issue with these data or my use of them, however, Vetta brings up other studies of the 1930s (presumably by N. Hirsch), none of which I have ever cited, involving IQ comparisons between various ethnic and national groups in the United States. I haven't looked into the merits or shortcomings of these old studies, but they apparently have a "bad image," which perhaps might attach to the data I have cited by N. Hirsch.

Such debating points impress on me the importance of urging that readers study *Bias* itself. What often appears as a trenchant criticism can turn out, when checked in the book, to be merely a restatement of my own caveats, or a denial of what was never asserted – or it may evaporate altogether when the point in question is read in its full context. A further reason that readers should consult the book itself is that some of the commentaries may leave a distorted impression of what the book is about; the discussions of heritability, and of the problems of cross-cultural testing, for example, are not at all central issues in this book.

But now to deal with the more substantive issues. I have organized my replies by topics, because there is considerable overlap among various commentators. This will facilitate dealing with the minor variations on the main themes.

Genetics and heritability of IQ and group differences

Bias emphasizes in its Preface that the study of test bias is the study of bias in the measurement of *phenotypes*. It is not an attempt to resolve the nature-nurture question. The one brief section of *Bias* (pp. 243–245) that mentions heritability is an attempt to substitute a better, more operational concept for the older notion of "capacity," and to explain the theoretical relationship between observed phenotypic measurements and hypothetical genotypic values in terms of the standard formulations of quantitative genetics. I have attempted to make this the clearest, most accurate, and most complete exposition of this point to be found anywhere in the psychological or genetic literature. The use of various kinship correlations in different populations as evidence of a test's construct validity (*Bias*, p. 427) has nothing to do with heritability per se, although it involves some of the same kinds of data that are used in the estimation of heritability.

The fact that the main arguments in *Bias* do not involve the concepts of genetics or heritability should not be misinterpreted to mean that I no longer believe that genetic factors are important in the abilities domain. It should be quite clear to everyone by now that one cannot properly speak of *the* heritability of IQ (or any other trait), although one can legitimately speak of the range and central tendency of the empirical *estimates* of heritability of a certain trait in a given population. My own position on this is essentially in agreement with the conclusions of the most recent reviewers of this growing body of evidence on the heritability of intelligence (e.g., Nichols 1979; Vernon 1979; Willerman 1979). The latest, most comprehensive review, emphasizing the most recent research, states, "Although we conclude that the new mental test data point to less genetic influence on IQ than do the older data, the new data nonetheless implicate genes as the major systematic force influencing the development of individual differences in IQ. In fact, we know of no specific environmental influences nor combinations of them that account for as much as 10 percent of the variance in IQ" (Plomin & DeFries 1980, pp. 21–22).

As to the question of the heritability of mean differences between racial populations, such as blacks and whites, I have stated my position most recently elsewhere (Jensen 1978) and I give a more extended treatment in my forthcoming book, *Straight Talk about Mental Tests* (Jensen 1981). In brief, I believe that the hypothesis of genetic differences between racial populations in some behavioral traits, including intelligence, is reasonable and plausible, but not validated by any method that would be acceptable to geneticists as rigorous direct support. Without a true genetic experiment involving cross-breeding of racial samples and the cross-fostering of the progeny of such random samples of every race × sex combination of the two populations in question, all other types of behavioral evidence can do no more than enhance the plausibility (or implausibility) of a genetic hypothesis. Whatever social importance one may accord to the race-genetics question regarding IQ, the problem is *scientifically* trivial, in the sense that the means for 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 necessarily taxes scientific ingenuity, and is hence probably insoluble.

This is a good place to point out the error in the often repeated cliché 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. The cliché is false. 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 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 heritability *within* groups increasingly exceeds zero, it 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. Yet this biggest straw man in the IQ controversy has been the chief target of many critics.

Humphreys's statement is the best reply to all those who have failed to heed the paragraph in my Preface (p. xi) stating that the heritability issue is not germane to my investigation of test bias. Humphreys emphasizes the additional important point (also made in *Bias*, p. 737) that the observed racial differences are real (i.e., not an artifact of the tests) and socially consequential regardless of whether the causes are due to genetic or environmental factors or some combination of the two. On these crucial points, Humphreys's essay should be read and reread.

Brody & Brody correctly note that twin correlations (and heritability estimates) in different groups can provide evidence of a test's construct validity in those groups. But there are many possible artifacts that can enter into group comparisons of heritability coefficients (and kinship correlations, although to a lesser extent). It is risky to compare heritabilities in different groups without taking account of restrictions of range, floor or ceiling effects, test reliability, degree of assortative mating in the parent population, and comparability of sampling. Certain kinships are also questionable in some groups and there must be controls for this. For example, in one of my studies I found a much higher frequency of half-siblings in the black than in the white population of Berkeley and had to take precautions against the possible contamination of full-sibling correlations by the greater admixture of misclassified half-siblings in the black sample. One would need to know at least the sample sizes and the variances for the twin correlations cited by the Brodys to evaluate the importance of the apparent social class difference. I have *hypothesized* lower IQ heritability for lower SES (socioeconomic status) groups (Jensen 1973, pp. 175–179), but establishing this empirically and interpreting it is no easy matter. One problem is the very large standard error of most heritability estimates, at least those based on the comparisons of MZ and DZ (mono- and dizygotic) twins, as in the correlations cited by the Brodys. There are so many less risky and less costly methods for detecting test bias and establishing construct validity that I would accord lower priority to the kinship methods explicated in *Bias* if one's only interest were in studying test bias. But I would also encourage the careful examination of kinship data collected for other purposes to assess their suitability for the detection of test bias.

Brace's notion that I have "disastrously used the concept of heritability" is entirely without foundation. Nor do I ever discuss black-white differences as if they were "eternal verities." (This is why I urge people to read the book itself.) It is merely a fact that white-black mean differences show up on every standard mental test whenever representative samples of each population are tested. Is it this fact that Brace refers to as an "eternal verity?"

If races actually differ genetically in intelligence or other socially important behavior-governing traits (and we do not know whether they do or do not), then the condition that **Brace** would require to make this determination may never be possible. It is precisely because the genotype becomes causally correlated with the environment and, in a manner of speaking, to some extent fashions its own environment that purely observational studies cannot settle the race-genetics question with respect to behavioral or even physical traits. A cross-breeding and cross-fostering experiment (possibly with *in vitro* fertilization) would answer the question.

Humphreys and **Gordon** have largely answered **Brace.** I will only add that the study of the accuracy and properties of the measuring instruments in any science is a worthwhile pursuit in its own right. That *Bias* attempted to do this in the case of mental tests, without encompassing every issue in the whole field of differential psychology, does not seem to me to be proper grounds for criticism. A division of the broad subject matter of differential psychology is a practical necessity. My previous writings and a forthcoming book (Jensen 1981) deal with other topics in differential psychology.

Hirsch claims that my formula 6.10 (*Bias*, p. 243) is "a thoroughly inappropriate application of a population parameter to the individual." He is wrong. It follows logically from the genetic model which formula 6.10 summarizes. Also it is conceptually not different from the estimation of true scores from obtained scores in classical test theory, with h^2 corresponding to r_{XX} (test reliability). It is merely the application of a simple regression equation. In quantitative genetics, the broad heritability h^2 can be conceived as the square of the Pearson correlation between genotypic and phenotypic values, and the genetic variance is $h^2\sigma_X^2$, where σ_X^2 is the phenotypic variance. It follows that the regression of genotypic values on phenotypic values is $h(h\sigma_X/\sigma_X)$ or h^2. The estimated genotypic value for any individual phenotypic value (P_i) in the population from which the parameters of the regression equation were derived in $\hat{G}_i = h^2 (P_i - \overline{P}_p) + \overline{P}_p$, where $\overline{P}_p$ is the phenotypic mean of the population.

Wahlsten's assertion that a genetic model such as $P = G + E$ is rendered invalid because hereditary and environmental factors interact throughout development represents a misun-

derstanding of the necessary distinction between two meanings of the term *interaction*. Those who disparage additive genetic models usually have in mind the kind of interaction of genotype and environment which merely amounts to the truism that every organism has a genotype and develops in an environment by interacting with it physically. Additive models, of course, do not in the least deny interaction in that obvious sense, nor does any model for partitioning the phenotypic variance. *Statistical interaction* of genotypic values and environmental values, on the other hand, is an empirical question, and for some traits (including IQ) additive genetic models without $G \times E$ interaction usually fit the data quite well. In fact, no one has yet been able to detect any significant component of IQ variance that is associated with $G \times E$ interaction. Conceptually, the simplest demonstration of such an interaction would be to compare two sets of MZ twins (say, twins AA′ and twins BB′), where A and B are both reared in environment X and A′ and B′ are both reared in environment Y. If, then, the trait difference (measured on an equal interval scale) between A and B is significantly greater than or less than the difference between A′ and B′, this would constitute evidence of $G \times E$ interaction.

Economos, in her opening paragraph, intimates that I used Burt's data to argue a point about the white-black IQ difference. The plausibility of the hypothesis that genetic as well as environmental factors are involved in the difference does not now, and never did, depend on Burt's data. The only time I have ever juxtaposed MZ twin differences and the mean white-black differences has been to point out the weakness of the argument of some environmentalist critics that the magnitude of MZ twin differences supported the hypothesis that the black-white difference in IQ is exclusively the result of environmental differences (Jensen 1973, pp. 161–173). Burt himself never showed any interest in the race question, either in his conversations with me or in his voluminous writings. The only thing on race differences I have ever come acrosss in Burt's writings is in a brief footnote in his study of Galton, in which he expresses some disagreement with Galton's belief in genetic racial differences in intelligence, particularly regarding blacks, and notes that population differences in variance may have more important cultural consequences than differences between the means (Burt 1962, p. 47).

Wahlsten thinks that a strict genetic or biological criterion of race, rather than the social definition, is important. But of course the ordinary social concept of race is the only appropriate one for the study of cultural bias in tests, for the reasons pointed out by Humphreys. Those who claim that, say, the SAT (Scholastic Aptitude Test) is biased against blacks are talking about people who are socially classified as blacks, who identify themselves as blacks, and who are perceived by others as blacks. Genetic criteria of African ancestry, such as the frequencies of certain genes, blood groups, skin reflectance, and various anthropometric indices are wholly irrelevant to the study of bias in tests for the practical purposes for which tests are generally used.

Vetta confuses my use of sibling regression with the heritability issue and the proof of a polygenic hypothesis, which is not at all central to this book. He does (correctly, I believe) point out the necessary relationship between a zero correlation between sibling means and differences (in test scores) and the normality of the distribution of the scores. The sibling method I have described, therefore, can be regarded as an indirect test of the normality of the distribution; and if we postulate that intelligence is normally distributed, the sibling method is then a test of interval scale for IQ (or other scores). Agreed, it is not an independent demonstration, but merely a logical corollary of the normal distribution. This valid point is unfortunately obscured by Vetta's intent either to disconfirm or to cast doubt on the heritability of IQ. But no one argues that the heritability of a given trait is of any particular value in general. Vetta has for a long time been a harsh critic of research on the genetics of intelligence and it is a pity he seems to have perceived *Bias* as simply one more opportunity to further his criticism of the heritability of IQ.

Osborne's mention of Urbach's (1974) interesting perspective on the heritability controversy from the viewpoint of the philosophy of science is very apropos.

Vogel is right, of course, in noting that the gene (and its mode of action) is now no longer a hypothetical construct but a biochemical fact. However, the concepts of genotype and genotypic value, as they are used in quantitative polygenic models of intelligence (or of any other polygenic traits) are still properly characterized as theoretical constructs. As Vogel says, the "black box" represented by these constructs will, we hope, gradually be broken up by future research at the interface between biometrical and molecular genetics.

General factor of cognitive ability

I will take some credit for helping to revitalize Spearman's concept of *g*, the general factor of cognitive abilities, although without necessarily endorsing any particular speculations Spearman offered concerning its basic nature. The overriding fact that must be reckoned with and ultimately understood is what Thurstone termed the "positive manifold" – the striking phenomenon of positive correlations among all tests of mental ability. Those types of factor or component analysis which extract a general factor (as a first principal factor or first principal component or as a higher-order factor extracted from the correlations among oblique first-order factors) can be viewed merely as a means of summarizing the fact of the positive manifold nature of any correlation matrix of mental tests, and of identifying those tests which share the most variance with all the others. Up to that point there would seem to be little to argue about. The meaning of the "best" *g* and the proper method for extracting *g* (in order to compare tests' factor loadings or to compute a person's *g* factor scores) are less settled matters. I have not found a unanimity of opinion on this issue among the world's leading experts on factor analysis, although if any one method of estimating the *g* of a correlation matrix is most favored, it would seem to be the first principal factor in a common factor analysis. But there are also good theoretical arguments for the hierarchical extraction of *g* as a higher-order factor, or two higher-order factors – fluid and crystalized *g*, à la **Cattell.** In practice, I find that the different methods of extracting *g* yield such similar results as to be practically equivalent in terms of any general conclusions one ordinarily draws from such an analysis. But I do wish the experts could tell us something more definitive on this score. I suspect that the question has no general analytic solution but could be answered satisfactorily for all practical purposes by Monte Carlo methods, in which artificially constructed populations of subjects and of tests (with completely "known" factorial structures) are randomly sampled and subjected to different methods of *g* extraction. The preferred method, I should think, would be the one which yields the smallest sampling variation in *g* factor scores. (Here is a possible doctoral dissertation for a graduate student with ample computer resources.)

Part of the problem in the field of factor analysis, I have come to believe, results from the fact that it has become largely the province of mathematical psychologists and statisticians with little or no interest in any substantive problem in psychology. In this, they are unlike all of the pioneers of factor analysis, who developed this method in pursuit of a psychological theory of abilities. Although different rotations of the factor axes are mathematically equivalent in accounting for the data, they are surely not equivalent in psychologi-

cal or theoretical interest. I urge the mathematical experts in this field to put their most esoteric projects on the back burners long enough to provide some new and possibly definitive methods for the logical discussion of the factor analytic aspects of *g*. Those of us who want to understand *g* at a deeper level than that of factor analysis need to know the best way to estimate *g*, so as to be able to pick out the best *g* marker tests and calculate the best *g* factor scores. This is needed for research both in experimental cognitive psychology and in behavioral genetics. The most important problem with *g*, of course, is not one of factor analysis at all, but one of scientifically explaining the empirical phenomenon of the positive manifold, which is merely summarized by *g*, and has confronted psychologists since Spearman first discovered *g* in 1904.

Vandenberg's comments on factor analysis are essentially in agreement with mine. Although I have emphasized the *g* factor because of its relevance to our concepts of intelligence, and also because it has been so neglected until quite recently, I recognize, with Vandenberg and **Tyler** the practical and theoretical importance of various non-*g* group factors, such as verbal, numerical, and spatial abilities. However, there has been a tendency to overrate these group factors, because the tests that measure them also measure *g*, and often more of *g* than of the group factors the tests are intended to measure. When the *g* variance is removed from such tests, as by the use of factor scores, the practical predictive validities of these non-*g* factors are usually surprisingly meager.

I agree with **Vandenberg** that no single IQ test is an ideal measure of *g*. IQ merely estimates *g*, which is a rather vaguely defined construct. And I agree with **Cattell** that Raven's matrices, although they are highly *g*-loaded, are less than ideal because of specific variance due to using only the matrix problem format. Cattell's Culture Fair Test of *g*, which employs several different types of nonverbal reasoning items, does not contaminate the *g* factor (actually fluid *g*) with variance specific to item type, as I have explained in *Bias* (p. 650).

The *g* factor can become more clearly defined only if we develop a more complete theory of it. As I have said clearly in *Bias* and elsewhere (Jensen 1979), and as **Kempthorne & Wolins** reiterate, *g* (or general intelligence) is not an entity or a "thing," but a hypothetical construct or a theory. It is the critics of intelligence testing who have tended to reify intelligence as a "thing," and not those who are actually doing research in this field.

I am not very confident of the method of rotation of factors after the extraction of *g*, mentioned by **Vandenberg,** although it has been done by a recognized expert in factor analysis, Maxwell (1972), with easily interpretable results. There may well be a better method, such as the Wheery method mentioned by **Kempthorne & Wolins.** But I am sure we cannot be satisfied merely with orthogonal rotation of the primary factors, which obscures the most salient feature of all correlation matrices in the abilities domain, namely the positive manifold. We need *g* and we need group factors for a sensible psychological picture. Of course, each of these factors can be subjected to further analysis by the nonfactor analytic techniques now being developed in experimental cognitive psychology [see Sternberg: "Sketch of a Componential Subtheory of Human Intelligence" *BBS* 3(4) 1980]. I see factors, and especially *g*, as the most interesting grist for experimental and psysiological analysis and theoretical explanation. Of course, the problem of factorial invariance, or the lack of it, has to be considered, as Vandenberg suggests. But I am struck by the relatively high degree of constancy of *g* across different test batteries (in the same population) and across different populations (for the same battery). This is an impressive empirical fact, which, to me, means we have a phenomenon eminently worthy of intensive scientific investigation. All hands on deck!

Sternberg indicates a misapprehension that I regard intelligence as the sum total of *all* abilities. I do not. I believe in the dimensional analyses (e.g., factor analysis) of the whole abilities domain; and I recognize the existence of a fair number of quite important non-*g* dimensions (group factors) and some undetermined number of smaller group factors, as well as considerable task specificity. I have never conceived of representing all abilities on a single dimension, as **Kempthorne & Wolins** suggest. But I recognize that practically every test has some positive *g* loading, whatever its loadings may be on other factors. The challenge is to try to find out why such seemingly disparate tests as block designs, vocabulary, backward digit span, and choice reaction time are all positively correlated with one another, even though the far from perfect correlations necessitate the hypothesis of other ability factors besides *g*. I would say the same about Kempthorne & Wolin's reference to artistic, musical, and cooking abilities. These, too, are *g*-loaded, although other non-*g* group factors are probably more important in such abilities than in most skills emphasized in school. Some threshold value of *g* seems to be necessary, but not sufficient, for the manifestation of other abilities or talents, even when these can be measured by tests with relatively small *g* loading. Spearman discovered that even pure pitch discrimination ability has some *g* loading when factor analyzed among a battery of cognitive tests. Certain highly *g*-loaded abilities, such as reading comprehension, are much more important societally than others. The fact that the basic scholastic skills are among the most important *g*-loaded types of performance, and are quite highly correlated with IQ tests, does not imply that IQ or other highly *g*-loaded tests predict *only* scholastic kinds of performance.

These remarks are pertinent also to **Eckberg**'s statements about *g*, which he says accounts for a relatively small amount of the full range of human achievement. But what other unitary factor accounts for more? I think *g* is the most important single factor in cognitive achievements of all kinds, which makes it extremely worthy of study, even though many other lesser factors (in the sense of variance accounted for) are involved. Whether or not *g* is a fruitful scientific construct must be decided on other than political grounds. To argue that some *g* theorists of the past (or present, for that matter) have espoused political views with which their critics disagree is wholly irrelevant to the scientific usefulness of the construct. It is in this *ad hominem* vein that **Hirsch et al** quote Spearman's (1914) statement relating *g* to eugenics, but the scientific importance of *g* clearly does not hinge on that point at all.

Hirsch et al's argument that *g* is "mythical" and "has no biological unity" because the genome fractionates is a confusion of different levels of analysis. The *g* factor emerges from the analysis of intercorrelations among various samples of molar behavior. The fact that many experimentally separable influences, including genetic factors, are involved in such molar behavior cannot contradict the *g* factor at this level of analysis. No one has ever thought of *g* as some kind of single, indivisible, irreducible "atom" within the brain. There is a large *g* of physical body measurements, just as of ability measurements, which can be characterized as a general body-size factor and which is obvious from just looking at the people around us. Would Hirsch et al say that it, too, has no "biological unity," whatever that might mean? Yet individual differences in body size must be every bit as polygenic as individual differences in mental ability, and the genomes are just as fractionable. I am not sure that Hirsch et al are going so far as trying to deny that genetic factors are involved in individual variation in *g*. Hardly anyone would bet on the prediction that *g* factor scores have zero heritability. To the extent that *g* is estimated by IQ, the answer, in general terms,

is already clearly established.

Brace refers to my statement that we do not yet have a true theory of *g* as if this somehow diminishes the importance of *g* or the reality of the empirical observation that led to the concept. There is still theoretical dispute in physics about the nature of gravitation, but that does not negate the various phenomena we recognize as examples of gravitation. When I see an empirical disconfirmation of the positive manifold in the abilities domain, I'll then rethink my great interest in *g*. **Vetta**'s suggestion that "some psychologists believe that scientific thought ended with Spearman's *g*" can be evaluated in the light of the foregoing discussion (and chapters 6 and 14 of *Bias*).

The Spearman hypothesis

Chapters 11 and 15 of *Bias* report several studies that support what I have termed the Spearman hypothesis – that the size of the white-black mean difference on a test is directly related to the test's *g* loading. I consider this an important, unifying hypothesis, which is supported by all the relevant data I have been able to find. Since the publication of *Bias*, four large new sets of relevant data have come to my attention (personal communications from: Lloyd G. Humphries; R. K. Osborne; Cecil Reynolds; Jonathan Sandoval), and all are consistent with the hypothesis, which I expect will become one of the most unequivocally substantiated hypotheses in differential psychology. So far I have found no data that contradict it.

Tyler and **Osborne** take note of this hypothesis and seem to appreciate its potential contribution to understanding the variation in magnitudes of the mean white-black differences in diverse tests and tasks. **Sternberg,** on the other hand, suggests that the hypothesis is not well supported. But I have now found ten studies altogether that are consistent with it (the four personal communications mentioned above, plus the Hennessey, Nichols, Osborne & Suddick, Veroff et al., U.S. Department of Labor, and Jensen studies cited in *Bias*), and none that are inconsistent; and I have sent out a letter to many psychologists who would be likely to know of relevant data sets in the hope of finding more evidence that would either confirm or refute the hypothesis.

Brody & Brody note that in one of my studies the Peabody Picture Vocabulary Test (PPVT) shows a slightly larger mean white-black difference than Raven's Progressive Matrices (RPM), and they believe this fact contradicts the Spearman hypothesis. I disagree, for two main reasons. First of all, the PPVT and RPM differ in reliability, and when the mean white-black differences in σ units are corrected for attenuation, the groups differ slightly more on the RPM than on the PPVT. Secondly, vocabulary is one of the most highly *g* loaded type of test; it measures crystalized *g*, while the Raven measures fluid *g*. The fluid eductive processes are involved in the original *acquisition* of vocabulary as explained in *Bias* (pp. 145–147). The PPVT and RPM are more highly contrasting tests in cultural loading than in *g* loading. Moreover, I believe that a proper test of the Spearman hypothesis should be based on the *g* loadings of tests derived from a factor analysis of at least several fairly diverse tests, rather than on a comparison of only two tests that may differ in a number of other psychometric features that make their interpretation ambiguous with respect to the Spearman hypothesis. Tests that differ in fewer respects than do the PPVT and RPM, such as tests of forward and backward digit span, clearly support the Spearman hypothesis. Backward digit span (BDS) is about twice as *g* loaded as forward digit span (FDS), and the mean white-black difference (in σ units) on BDS is about twice as great as on FDS.

Average white-black IQ difference

Gordon reviews the evidence for **Cattell**'s observation that the average white-black IQ difference has remained fairly constant for a great many years. **Osborne** cites evidence, released by the Educational Testing Service after *Bias* went to press, showing white-black differences at the graduate level, on the Graduate Record Exam, the Law School Admissions Test, Medical College Admissions Test, and other high-level scholastic aptitude tests. These white-black mean differences, in σ units, are as large as or larger than the overall population differences found on IQ tests, but this finding should be viewed in relation to the highly restricted σ of test scores in the segment of the population that seeks admission to graduate and professional schools – the differences on an IQ scale may be considerably less than would appear in terms of the sigmas of this highly self-selected sample.

Humphreys notes the lower black mean IQ at the *end* of the public school period. But it should not be overlooked that the black IQ deficit is just as great at the age of entering school, except in the rural South, where there has been some indication of a declining black IQ as children advance in age. Thus, the schools, by and large, cannot be blamed for creating the black IQ deficit, which is fully evident at school entry; but neither do the schools at present do anything to ameliorate it.

Vetta proposes a better way to test the differences between the ratios of black-white means and standard deviations, by using the standard error of the coefficient of variation, on p. 425 of *Bias*. I accept this solution. However, it should not be misconstrued as an argument against the basic logic of comparing black and white regressions of raw scores on age, which must surely take account of the group difference in mental growth rates, by which I simply mean the blacks' slower rate of increase, relative to whites, in average mental age scores (or raw scores) as a function of chronological age.

Economos has the mistaken impression that I rely on socioeconomic status (SES) indices "to subtract blackness from test results." What I do demonstrate, however, is that the white-balck differnce in not just an SES difference by any generally accepted index of SES. SES group differences (within each race) show a different pattern or profile across various tests than is shown by white-black differences. The important point is that the evidence clearly shows that the white-black difference is not just an SES difference, whatever else it might be. In any case, I do not view SES as mainly a causal variable in relation to IQ; to do so is what I have termed "the sociologist's fallacy," (Jensen 1973, p. 235).

IQ and the normal distribution

Whatever may be its shortcomings, I believe that the discussion of the distribution of mental ability in chapter 4 of *Bias* is probably better than any other treatment of this topic in the psychological literature. No doubt a better treatment is possible, and I hope someone better qualified than I will put forth the effort.

Wahlsten's comments on the distribution of IQ should be compared with what I actually claimed, which is that IQ is *not* normally distributed, but shows marked departures from normality at the lower and upper extremes. IQ does, however, conform very closely to the normal curve in the central region between about $\pm 2.5\sigma$ from the mean. This was also the main point of Burt's two articles on the distribution of IQ! I am amazed at Wahlsten's notion that Burt would have "faked" a normal distribution of IQ in order to support his conclusion that IQ is *not* normally distributed (but resembles a Pearson Type IV curve).

Wahlsten's and **Dorfman**'s censure of my references to

Burt in this context are examples of the poison-well fallacy. Some of Burt's purported data on MZ twins reared apart are very probably fraudulent. But should that rule our reference to his discussion of the form of the IQ distribution, which is probably the best theoretical treatment of the issues I have found in the literature? The IQ data he presents from the schools of greater London show essentially the same distributional features as the normative data on the Stanford-Binet and Wechsler IQ tests in the United States. My inference that environmental effects on IQ are normally distributed does not depend on Burt's data on MZ twins reared apart, as Dorfman's comment might lead one to believe. Data from three other studies of MZ twins reared apart, totalling 69 twin pairs (versus Burt's 53 pairs) show the same distributions of absolute differences in IQ between twins as did Burt's partially fictitious twin data. I had already looked into that point long ago (Jenson 1974b).

I agree with **Dorfman** and with **Kempthorne & Wolins** that the form of the distribution of achievements cannot be argued on the basis of ordinary achievement tests. But I gave a rationale for hypothesizing that the distribution of achievements would be skewed and pointed out the interesting fact that types of achievements that can be enumerated, that is, that have a true zero point and can be discretely counted (vocabulary size, number of patents of inventors, number of publications of professors, and the like), show markedly skewed distributions. I would not hastily dismiss this interesting finding; we should try to understand its relevance to the distribution of abilities, which, when measured on a true scale, generally seem more closely to approximate a normal distribution than does the skewed distribution of enumerable achievements.

If **Vetta** or anyone else objects to a distribution *approximating* the normal curve as a reasonable assumption about the distribution of ability, he should propose and defend some other assumption that may be more suitable. Assumptions and boundary conditions are necessary tools in all sciences. They are to be evaluated in terms of their utility in theory and application. The theory that intelligence is approximately normally distributed (within the limits of about $\pm 2.5\sigma$ of the mean) is no exception.

Meaning of bias

Kempthorne & Wolins say they do not find "a reasonable general definition of test bias" but agree with my ideas and procedures for examining test bias. I defined bias as "systematic errors in the *predictive validity* or the *construct validity* of test scores of individuals that are associated with the individual's group membership" (*Bias*, p. 375). I later explicate various types of bias and psychometric manifestations of bias.

Sternberg contrasts what he refers to as my "narrow" treatment of test bias with some other kind of test bias that is presumably more "broad" but less clearly defined. I suspect that underlying this criticism is an implicit assumption that our mental tests ought to provide accurate measurements of genotypic rather than phenotypic values. **Kempthorne & Wolins** state what I think is the correct response to this notion, in their height-weight analogy. Actually, what we want our test scores to assess accurately is the phenotype. The fact that phenotypes may change in time or be altered by external influences should be reflected by test scores. Tests should be judged by how well they measure what *is* rather than what we may think ought to be or what might be in a different world. We do not say yardsticks are biased because children who have survived a famine are undersized for their age; nor would the measuring device be as useful if it reflected, not phenotypic size, but what the child's size might have been under some optimal conditions (not necessarily equal for everyone) for the development of size for each child's particular genotype. Despite Sternberg's complaint, my conclusion really isn't so "narrow" after all. What I have shown is that one popularly supposed source of racial and social group differences – the tests themselves – does *not* contribute to the differences. In this sense, as **Humphreys** points out, the differences, and their societal correlates, are real. They are not artifacts of the tests or the test situation. Moreover, we can only look for bias in the tests that already exist. If other potentially important abilities have not yet been measured, then we should devise tests to measure them, if possible, and demonstrate their practical validity. They, too, should be examined for bias, as I have done with many existing tests.

I agree with **Sternberg** that tests may discriminate against individuals who experience test-taking difficulties, anxiety, and the like. This is not unrecognized by psychometricians and is one of the many factors contributing to the far less-than-perfect validity of tests. (Considerable efforts are being made to take account of such personality factors in testing.) But such sources of unwanted variance do not nullify the part of the test variance that *does* contribute to their validity. Unless we see both factors in this perspective, we unwittingly reinforce those who see nothing at all good in tests and whose main goal is the complete abolition of all tests.

Green predicts that latent trait theory will become the chief method for analyzing test bias. I agree, at least for the detection of item bias. I explained in the Preface (p. xi) why I did no more than introduce latent trait concepts (*Bias*, pp. 442–445 and 580), mainly because it has been so untried in this field as yet. Lorrie Shepherd (University of Colorado; personal communication) had recently compared latent trait methods for detecting item bias with many of the other methods employed in *Bias* and finds considerable agreement among the methods. Latent trait methods, however, require enormous samples. Green is in an excellent position to apply these techniques to the massive achievement test data obtained by the California Test Bureau, and we would all be grateful for a comprehensive monograph on bias in achievement testing.

Criterion validity of tests

Vogel notes the interesting fact that our current IQ tests are capable of reflecting quite subtle effects which are definitely attributable to genetic conditions. Such tests are used in assessing response to the dietary treament of PKU (phenylketonuria). It has also been found that individuals who are heterozygous carriers of the recessive PKU gene have lower IQs, by an average of about 10 points, than their noncarrier full siblings (Bessman, Williamson, & Koch 1978). IQ is also correlated with myopia (*Bias*, p. 362). In light of these and many other recent findings, it is increasingly apparent that IQ tests do to some extent reflect biological realities and not merely prior learning of skills and knowledge, as some prominent psychologists still seem to belive (e.g., Albee 1980).

To **Brody & Brody I** can only repeat what I said in *Bias* (ch. 8): IQ correlates much more highly with job status than with ratings of job performance *within* occupations, because of restriction of range and reliability of the criterion. The societal problems of minorities, particularly blacks, where job selection by tests is concerned, is in job status, rather than performance within jobs. These status differences mainly reflect the average white-black difference in IQ distributions and the relatively high correlation between job status and IQ,

independently of race. I have long agreed with the Brodys' suggestion that psychologists and educators should try to develop job-training methods that will somewhat reduce the dependence of their success on *g*. I have seen successful examples of this in some armed-forces training programs and in sheltered workshops for the mildly retarded.

Harrington objects to the correction of predictor variables for attenuation, as by the use of estimated true scores. *Bias* is thoroughly explicit and detailed on this point, as noted by **Kempthorne & Wolins.** There is no basis for Harrinton's objection. It can be shown logically and empirically that predictions of a criterion from test scores in two (or more) groups can often be made more accurately from estimated true scores than from unadjusted scores. Many instances of apparent predictive bias, usually manifested as intercept bias, are eliminated by correcting the predictor variables for attenuation, that is, using estimated true scores as the predictor variable. Harrington's assertion that "If a test is biased we can't estimate the true score" is evidently based on a misconception of the meaning of "true score." If one knows the mean and reliability of the test scores for the groups in question, estimated true scores can always be obtained, and they are, by definition (and by empirical fact), less biased than the uncorrected scores.

One can argue, of course, that the test and the criterion are equally biased, and therefore test bias escapes detection by some of the methods I have described. I believe that the extremely wide variety of performance criteria in schools, colleges, armed forces training programs, and dozens of different job categories for which aptitude tests show no significant bias in predictive validity for blacks and whites renders it extremely implausible that the same degree of bias exists in the performance criteria as in the tests. In this case, what may be *theoretically possible* is in reality highly implausible.

Economos refers to *Who Gets Ahead* by Jencks et al. (1979) as evidence that IQ has little relevance for occupation and income level in our present society. But, from the very same data base, Eckland (1980) has argued – more correctly, I believe – that a quite contrary conclusion is actually warranted. The issues are too complex for discussion here; interested readers should consult Eckland's important article. On this point, **Eckland**'s accompanying comments are also a reply to Economos.

Eckland's discussion of the importance of the IQ of teachers is one more impressive item of evidence for the practical validity of IQ, and it underscores the value of *g*-loaded tests, such as vocabulary, for the selection of teachers.

Longstreth, in addition to presenting a good summary of the main conclusions of *Bias*, cites new evidence (Messé et al. 1979) that standard ability scores have had correlations of .60 with teachers' ratings of pupils' scholastic performance but are less SES biased than teachers' ratings, which tend to overpredict the performance of low SES pupils. Other studies have shown that IQ has lower predictive validity for teacher-assigned grades than for scores on achievement tests, but this is explained by the lower reliability of grades and the heterogeneity of the criterion, when composite grades include performance in subjects as diverse as physical education, art, and music, which do not reflect *g* as much as they reflect other factors (e.g., Henderson, Butler, & Goffeney 1969; Goldman & Hartig 1976). WISC-R IQ correlates with teachers' ratings of pupils in reading and mathematics even more highly for black than for white pupils (Reschly & Reschly 1979). I am not aware that anyone has ever presented a good case for the proposition that teachers' ratings and grades are a more reliable or more valid or less biased indicator of scholastic achievement than standardized achievement tests.

Breland notes other studies of the SAT not cited in *Bias* that lend further support to the "no bias" conclusion; I recommend his comprehensive review of bias studies of the SAT (Breland 1979), which came out after *Bias* was in press but fully confirms its conclusions about the SAT with respect to predictive bias or the lack of it. (*Bias* does discuss certain problems and studies which Breland [4th paragraph] claims are overlooked; I refer him to pages 331–332, 329, and 467–468.)

Reynolds, too, cites new evidence that is fully consistent with the conclusions of *Bias* and provides a crucial extension of the conclusions to the validity of IQ tests for children placed in special classes, such as the educable mentally retarded – the issue that figured prominently in the Larry P. trial in California.

Kempthorne & Wolins seem to disapprove of the exposition of intelligences A, B, and C (*Bias*, pp. 183–184), which are attributable to Hebb (1949) and Vernon (1979), whose conceptually useful formulations, I believe, are accurately summarized. If intelligences A, B, C are not regarded as conceptually useful distinctions, I am sure many psychologists would like to know the reason why.

Animal intelligence

Wahlsten's and **Hirsch et al.**'s excursions into the details of zoological taxonomy, however correct they may be, are irrelevant to any of the main points in my discussion of animal intelligence. This criticism seems a good deal like arguing about which is correct, "two plus two *is* four" or "two plus two *are* four," when what one is interested in the arithmetic and not the grammar. I was not concerned with whether or not *all* species can be ordered on a single dimension of intelligence. My point was that people have no trouble at the commonsense level in recognizing differences in intelligence between, say, chickens, dogs, and chimpanzees, and I pointed out the features of these animals' typical behaviors in experimental situations that are most discriminating with respect to our impressions of their rank order in what we commonly regard as "intelligence." It is most interesting that these are also the very features that best characterize *g* loaded test items for humans, and performance on certain animal tests taken by human children and the mentally retarded is related to their intelligence levels just as is performance on the Stanford-Binet. Briefly, *reasoning*, more than *learning*, characterizes our notion of intelligence, both among animal species and among humans. These parallels between our conceptions of animal and human intelligence are theoretically interesting, regardless of the taxonomic and evolutionary relationships between the various species studied by comparative psychologists.

Hirsch et al. apparently believe that Harrington's (1975) study of six genetic strains of rats, which learned various mazes with scorable units analogous to items in psychometric tests for humans, supports the generalization to humans that "Majorities will score higher than minorities as a general artifact of test construction procedures." Harrington's clever study is of great interest in its own right. But should we not question the generalizability of this rat experiment to the psychometric testing of humans, when, in fact, the human evidence flatly contradicts this generalization with respect to actual IQ tests? Asians and Jews are minorities in America that score as high as or higher than the majority on majority-standardized IQ tests. Japanese in Japan, on the average, out perform 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. African infants consistently score more highly on the American-standardized Bayley Infant Scales of Development than do middle-class white American infants for whom this test was originally

developed. Many other counterexamples to Harrington's generalization could be cited. What Hirsch et al. should be asking is why Harrington's finding, based on genetically different strains of rats (a fascinating finding, incidentally), is so thoroughly at odds with test data on different human populations. I hope someone will pursue this question, which I see as being of greater relevance to behavioral genetics than to human psychometrics.

Pygmalion and IQ

Rosenthal's comment reinforces my conclusion (*Bias*, p. 608) that teacher expectancy (or the so-callled "Pygmalion effect") has still not been demonstrated with respect to IQ. The fact that I inadvertently included three studies involving scholastic achievement rather than IQ tests in my list of thirteen studies that have failed to show a significant expectancy effect does not, of course, contradict the fact that nine studies performed since the questionable original Pygmalion study by Rosenthal and Jacobson (1968) afford no substantiation of the expectancy effect for IQ, and no study has shown the effect for IQ. Scholastic achievement would seem to be more susceptible to a teacher-expectancy effect, but the studies of it, too, are unimpressive. A tabulation of other studies showing significant expectancy effects for *other* dependent variables, no matter how extensive, of course, cannot weaken the overwhelmingly negative conclusion with respect to IQ tests.

What I consider the most powerful study ($N = 416$) of examiner expectancy on WISC Performance IQ (Samuel 1977) indeed showed 6.4 percent of the total variance associated with 31 sources of variance in the ANOVA (main effect and all 30 interaction terms involving expectancy), but neither the main effect nor any of the first- or second-order interactions were statistically significant. Considering the lack of significance and the fact that no variance component can be less than zero, the fact that 31 terms associated with expectancy sum up to only 6.4 percent of the total variance surely cannot be construed as support for the Pygmalion effect on IQ. It all adds up to a trivial or nonexistent effect. It is grasping at straws to claim the contrary.

Statistical matters

Dorfman takes issue with my definitions of path analysis and correlation. But the reader can compare the three quotations from *Bias* that are criticized by Dorfman with the following statements from well-known textbooks on path analysis and statistics:

> "... path analysis is concerned with erecting a causal structure compatible with the observed data" (Li 1975, p. 3).
>
> And from the very first paper on path analysis by its inventor: " The present paper is an attempt to present a method of measuring the direct influence along each separate path in such a system and thus finding the degree to which variation of a given effect is determined by each particular cause. The method [of path analysis] itself depends on the combination of knowledge of degrees of correlation among the variables in a system with such knowledge as may be possessed of the causal relations" (Wright 1921).
>
> "A *coefficient of correlation* is a single number that tells us to what extent two things are related, to what extent variations in the one go with variations in the other" (Guildford 1956, p. 135).
>
> "... the square of the correlation coefficient gives the proportion of the total variance of Y which is predictable from X, or r^2 measures the proportion of the Y variance which can be attributed to variation in X" (McNemar 1949, p. 116).

Although **Kempthorne & Wolin** express complete agreement with the main conclusions of *Bias*, there seems to me to be a good deal of hyperbole in their repeated assertions as to "mistakes," "errors," "inconsistencies" of a statistical and methodological nature, without clearly specifying what these are. Fortunately, they have sent to me a much more detailed critique, over eight times the length of their published version, which lists all the "errors" that they claim. This longer critique [to appear in Continuing Commentary in a forthcoming issue of *BBS*] surely testifies to Kempthorne & Wolins's close and careful scrutiny of every page of *Bias* (or at least those that contain any statistical material), and in many respects I find their detailed criticisms – harsh though honest – the most useful of all. The specific listing of what they consider to be errors is actually much less damaging than the rather sweeping and damning statements they make in the shorter review co-published here. As one example, they write that chapter 7 "contains many other statistical and conceptual errors." Their detailed critique, however, mentions only three errors in this chapter. One of these is not really an error at all, but their confusion of the formulas for (*a*) the standard error of a difference between obtained scores of two individuals on the same test and (*b*) the standard error of the difference between two scores on different tests obtained by the same individual; *a* and *b* require two distinct formulas (p. 294 and p. 684) and each is correct as presented. The other two "errors" are points that are still being debated in the literature, for example, the use of the *F* test in an analysis of variance based on dichotomous (0 or 1) data, such as test items scored right or wrong. Some authorities have argued that the *F* test is sufficiently "robust" for this not to be a serious concern, and the ANOVA of item matrices is a common practice.

Altogether there seem to be seventeen "errors" listed in the long version of the critique by **Kempthorne & Wolins.** At least three of these are clearly statistical errors which can be corrected in the next printing of *Bias;* they are not critical for any of the book's conclusions. Three do not seem to me to be errors at all. The remaining eleven are doubtful or arguable points: some of these are not agreed upon by other experts in the field, others are generally unresolved problems, others again are problems of the underlying mathematical models of certain statistics, which will have to be resolved by experts in mathematical statistics. Most of the "errors" of the kind Kempthorne & Wolins are referring to are not of the clear-cut, all-or-none variety, but range along a continuum of problematic points, disagreements, and shades of mathematical rigor. It will be most valuable, therefore, when their detailed critique, which I find generally admirable, even if at points its style is unduly hyperbolic, appears in continuing commentary so that those who wish to can delve into these statistical matters. Not all the basic statistical problems of psychometrics have yet been resolved. To discuss fully each of these points raised by Kempthorne & Wolins would be impossible here.

One peculiarity in **Kempthorne & Wolins**'s style of which readers should be aware is their tendency to make statements in such a way that they could easily be misinterpreted as points of disagreement with me when they are in fact simply paraphrases of my own statements or of statements by others with which I, too, disagree, such as the phrase "innate learning abilities" as used by Judge J. Skelly Wright in the Hobson vs. Hansen decision.

Harrington is critical of the Group Difference/Interaction (*GD/I*) ratio that I proposed (*Bias*, pp. 561–562) as one index of item bias. This ratio of the group difference to the groups × items interaction (both scaled in terms of the

average individual differences variance within groups) is simply a means of summarizing in a single quantitative index the key results of a groups × items ANOVA. It permits the ordering of different tests on a scale representing the magnitude of the mean difference between groups relative to the groups × items interaction, which is an indicator of item bias. The smaller the *GD/I* index, the greater is the likelihood that a different selection of test items from the same population of items would eliminate or reverse the mean group difference. We find this to be the case for sex differences on IQ tests, for example, while it is never the case for white-black differences. If anyone can suggest a better index than GD/I for expressing this property of a test with respect to two (or more) populations, it would be welcomed.

Cross-cultural testing

Kline is concerned with the problems of cross-cultural testing. These I consider to be effectively answered with respect to the groups considered in *Bias*, namely, American-born, English-speaking minorities in the United States. To say that current standardized tests cannot be used on these populations (in which they are found not to be biased predictors) because such tests might be biased when used with remote cultural groups such as Eskimos and Bushmen is not unlike arguing that a clinical thermometer is problematic because it can't measure the temperature in a blast furnace. For those who are interested in approaches to remote cross-cultural testing, where educational background, language, customs, and values differ tremendously, *Bias* (p. 636) provides an ample list of key references.

Implications

Tyler correctly observes that I have not attempted to range very far in my discussion of the broader societal implications of the main findings of *Bias*, although the general tenor of my philosophy about the treatment of individual and group differences is indicated in the final chapter. It was well beyond the intended scope of *Bias* to deal with the causes and possible remedies for subpopulation differences. Notions about eliminating group differences cannot overlook a possible clash with our democratic notions of human freedom, and social and educational programs aimed at reducing group differences in certain traits seem to imply much greater burdens and restrictions of personal freedom for members of some groups than for others. Surely these problems will need to be aired, but neither *Bias* nor this Response would be the proper forum. The beginnings of such a discussion, however, are most clearly evident in the accompanying commentaries by **Cattell, Clarke, Eckland, Gordon,** and **Havender,** whose contrast of collectivist and individualist philosophies of social justice will most probably be pivotal in the debate about the broader implications of the main conclusion of *Bias in Mental Testing.*

References

Angoff, W. H. (1975) The investigation of test bias in the absence of an outside criterion. Paper presented at the National Institute of Education Conference on Test Bias, Washington, D.C. [DRG]

Albee, G. W. (1980) Open letter in response to D. O. Hebb. *American Psychologist* 35:386–387. [ARJ]

Bessman, S. P., Williamson, M. L., & Koch, R. (1978) Diet, genetics and mental retardation interaction between phenylketonuric heterozygous mother and fetus to produce nonspecific diminution in IQ: Evidence in support of the justification hypothesis. *Proceedings of the National Academy of Sciences* (USA) 75:1562–1566. [ARJ]

Block, N. J., & Dworkin, G., eds. (1976) *The IQ controversy*, Part II: Genetic component of IQ differences. New York: Pantheon Books/Random House. [JE]

Bloom, B. S. (1976) *Human characteristics and school learning.* New York: McGraw-Hill. [NB]

Board of Education. (1939) Report of the consultative committee on secondary education with special reference to grammar schools and technical schools. London: H.M. Stationery Office. [DDD]

Bock, R. D., & Vandenberg, S. G. (1968) Components of heritable variation in mental test scores. In: *Progress in human behavior genetics*, pp. 1–2, ed. S. G. Vandenberg. Baltimore, Md.: Johns Hopkins University Press. 1968. [SGV]

Bodmer, W. F. & Cavalli-Sforza, L. L. (1970) Intelligence and race. *Scientific American* 223:19–29. [DW]

Borgen, F. H. (1972) Differential expectations? Predicting grades for black students in five types of colleges. *Measurement and Evaluation in Guidance 4 (4):206–212.* [HMB]

Bowles, S., & Gintis, H. (1976) *Schooling in capitalist America.* New York: Basic Books. [DLE]

Boykin, A. W. (in press) Experimental psychology from a black perspective: Issues and examples. *Journal of Black Studies.* [RJS]

Brace, C. L. (1971) Introduction to Jensenism. In: *Race and intelligence*, ed. C. L. Brace, G. R. Gamble, & J. T. Bond, Anthropological Studies No. 8, pp. 4–9. Washington, D.C.: American Anthropological Association. [CLB]

Breland, H. M. (1977) *Group comparisons for the Test of Standard Written English*, College Board Research and Development Report (RDR-77-78, No. 1). Princeton, N.J.: Educational Testing Service. [HMB]

(1979) *Population validity and college entrance measures.* New York: College Board. [HMB, BKE, ART]

Brigham, C. C. (1923) *A study of American intelligence.* Princeton, N.J.: Princeton University Press. [RAG]

Brody, E. G., & Brody, N. (1976) *Intelligence: Nature, determinants and consequences.* New York: Academic Press. [NB]

Brown, A. L., & French, L. A. (1979) The zone of potential development: Implications for intelligence testing in the year 2000. In: *Human intelligence: Perspectives on its theory and measurement*, ed. R. J. Sternberg & D. K. Detterman, Norwood, N.J.: Ablex. [RJS]

Brown, A. W. (1930) Change in intelligence quotients in behaviour problem children. *Journal of Educational Psychology* 21:341–50. [AV]

Burks, B. S. (1928) The relative influence of nature and nuture upon mental development: A comparative study of foster parent-foster child resemblance and true parent-true child resemblance. *Yearbook of the National Society for the Study of Education* 27:219–316. [RAG]

Burt, C. (1925) *The young delinquent.* London: University of London Press. [DDD, ARJ]

(1962) Francis Galton and his contributions to psychology. *British Journal of Statistical Psychology* 15:1–49. [ARJ]

Callaway, E. (1975) *Brain electrical potentials and individual psychological differences.* New York: Grune & Stratton. [ARJ]

Cattell, R. B. (1971) *Abilities: Their structure, growth and action.* Boston: Houghton Mifflin. [RBC, PK]

(1972) Beyondism: A morality from science. New York: Pergamon Press. [RBC]

Centra, J. A., Linn, R. L. & Parry, M. E. (1970) Academic growth in predominantly Negro and predominantly white colleges. *American Educational Research Journal* 1:83–98. [HMB]

Clarke, A. M., & Clarke, A. D. B. (1979) The cardinal sin. *Nature* 282:150–151. [DDD]

Clear, T. A., Humphreys, L. G., Kendrick, S. A., & Wesman, A. (1975) Educational uses of tests with disadvantaged students. *American Psychologist* 30:15–41. [ARJ]

Cohen, J. (1977) *Statistical power analysis for the behavioral sciences.* Rev. ed. New York: Academic Press. [RR]

Cole, J., Gay, J., Glick, J., & Sharp, D. (1971) *The cultural context of learning and thinking.* New York: Basic Books. [RJS]

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966) *Equality of educational opportunity.* Washington, D. C.: Government Printing Office. [BKE, RAG]

Crano, W. D., Kenny, J., & Campbell, D. T. (9172) Does intelligence cause achievement? A cross-lagged panel analysis. *Journal of Educational Psychology* 63:258–275. [NB]

Cronbach, L. J. (1975) Five decades of public controversy over mental testing. *American Psychologist* 30:1–14. [RAG]

Crow, J. F. (1969) Genetic theories and influences. *Harvard Education Review* 39:301–309. [AV]

Darlington, R. B., Royce, J. M., Snipper, A. S., Murray, H. W., and Lazar, I. (1980) Preschool programs and later school competence of children from low-income families. *Science* 208:202–204. [JH]

Darwin, C. (1859) *On the origin of species by means of natural selection.* London: John Murray. [CLB]

DeFries, J. C., Kuse, A. R., & Vandenberg, S. G. (1979) Genetic correlations, environmental correlations, and behavior. In: *Theoretical advances in behavior genetics*, ed. J. R. Royce, pp. 398–417, Alphen aan den Rijn, The Netherlands: Sitjthoff & Noordhoff. [SGV]

Dietz, S. M. & Purkey, W. W. (1969) Teacher expectation of performance based on race of student. *Psychological Reports* 24:694. [RR]

Department of the Army. (1976) *Staff officer's field manual: Organizational, technical and logistic data. Field manual 101-10-1* Baltimore: U.S. Army AG Publications Center. [RAG]

DHEW (Department of Health, Education, and Welfare). (1976) Intellectual development and school achievement of youths 12–17 years: Demographic and socioeconomic factors. *Vital and health statistics: Series II, Data from the National Health Survey; no. 158. DHEW publication; no. (HRA) 77-1640.* Washington, D.C.: U.S. Government Printing Office. [RAG]

Dorfman, D. D. (1978) The Cyril Burt question: New findings. *Science* 201:1177–1186. [DW]

Dunnette, M. D. (1964) Critics of psychological tests: Basic Assumptions: How Good. *Psychology in the Schools* 1:63–9. [JH]

Dusek, J. B. & O'Connell, E. J. (1973) Teacher expectancy effects on the achievement test performance of elementary school children. *Journal of Educational Psychology* 65:371–377. [RR]

Eckberg, D. L. (1979) *Intelligence and race: The origins and dimensions of the IQ controversy.* New York: Praeger. [DLE]

Eckland, B. K. (1980) Education in the meritocracy. *American Journal of Education* 88 (August) [ARJ]

Eckland, B. K., & Alexander, K. (1980) The national longitudinal study of the high school senior class of 1972, In: *Longitudinal perspectives on educational attainment*, ed. A. C. Kerckhoff, pp. 189–222. Greenwich, Conn.: JAI Press. [BKE]

Elashoff, J. D., & Snow, R. E. (1971) *"Pygmalion" reconsidered.* Worthington, Ohio: Charles A. Jones. [RR]

Epps, E. G. (1978) Effects of testing conditions on performance of minority children. In: *Achievement testing of disadvantaged and minority students for educational program evaluation*, ed. M. J. Wargo & D. R. Green. Monterey, Calif.: CTB/McGraw-Hill. [DRG]

ETS (Educational Testing Service). (1979) Statement submitted to U.S. House of Representatives Subcommittee on Civil Service. May 15. [RTO]

Farver, A. S., Sedlacek, W. E., & Brooks, G. C. (1975) Longitudinal predictions of university grades for blacks and whites. *Measurement and Evaluation in Guidance* 7 (4):243–250. [HMB]

Feller, W. (1957) *An introduction to probability theory and its applications*, vol. 1. New York: Wiley. [DDD]

Feuerstein, R. (1979a) *The dynamic assessment of retarded performers: The learning potential assessment device, theory, instruments, and techniques.* Baltimore, Md.: University Park Press. [RJS]

(1979b) *Instrumental enrichment: An intervention program for cognitive modifiability.* Baltimore, Md.: University Park Press. [RJS]

Fienberg, S. E. (1977) *The analysis of cross-classified categorical data.* Cambridge: MIT Press. [DDD]

Fischbein, S. (1980) IQ and social class. *Intelligence* 4:51–63. [NB]

Fisher, R. A. (1918) Correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society (Edinburgh)* 52:399–433. [AV]

Fulker, D. W. (1973) A biometrical genetic approach to intelligence and schizophrenia. *Social Biology* 20:266–275. [RBC]

Fuller, J. L., & Thompson, W. R. (1978) *Foundations of behavior genetics.* St. Louis: Mosby. [DW]

Galton, F. (1892) *Hereditary genius.* 2nd ed. Cleveland: World. [DLE]

Goldberger, A. S. (1979) Heritability. *Economica* 46:327–347. [JH]

Goldman, R. D. & Hartig, L. K. (1976) The WISC may not be a valid prediction of school performance for primary grade children. *American Journal of Mental Deficiency* 80:583–587. [ARJ]

Goldman, R. D., & Hewitt, B. N. (1976) Predicting the success of black, Chicano, Oriental, and white college students. *Journal of Educational Measurement* 13 (2):109–117. [HMB]

Goldman, R. D., & Widawski, M. H. (1976) A within subjects technique for comparing college grading standards: Implications in the validity of the evaluation of college achievement. *Educational and Psychological Measurement* 36 (2):381–390. [HMB]

Goodnow, J. J. (1976) The nature of intelligent behavior: Questions raised by cross-cultural studies. In: *The nature of intelligence*, ed. L. B. Resnick Hillsdale, N.J.: Erlbaum. [RJS]

Gordon, E. (in press) *Human diversity and pedagogy.* Westport, Conn.: Mediax Press. [RJS]

Gordon, R. A. (1976) Prevalence: The rare datum in delinquency measurement and its implications for the theory of delinquency. In: *The juvenile justice system*, ed. M. W. Klein. Beverly Hills, Calif.: Sage. [RAG]

(in press, a) Labelling theory, mental retardation, and public policy: *Larry P.* and other developments since 1974. In: *The labelling of deviance: Evaluating a perspective*, ed. W. R. Gove. Beverly Hills, Calif.: Sage. [RAG]

(in press, b) Research on race, IQ, and delinquency: Taboo or not taboo. . . ? In: *Taboos in criminology*, ed. E. Sagarin. Beverly Hills, Calif.: Sage. [RAG]

Gordon, R. A., & Rudert, E. E. (1979) Bad news concerning IQ tests. *Sociology of Education* 52:174–190. [ARJ]

Gould, S. J. (1980) Jensen's last stand. *New York Review of Books* 27(7):38–44. [GMH, JH, OK]

Gozali, J., & Meyen, E. L. (1970) The influence of the teacher expectancy phenomenon on the academic performances of educable mentally retarded pupils in special classes. *Journal of Special Education* 4:417–424. [RR]

Green, D. R. (1975) Procedures of assessing bias in achievement tests. Paper presented at the National Institute of Education Conference on Test Bias, Washington, D.C. [DRG]

(1976) Redicing bias in achievement tests. Paper presented at the meeting of the American Educational Research Association, San Francisco. [DRG]

Guilford, J. P. (1954) *Psychometric methods.* 2nd ed. New York: McGraw-Hill. [GMH]

(1965) *Fundamental statisitcs in psychology and education.* 3rd ed. New York: McGraw-Hill. [ARJ]

(1967) *The nature of human intelligence.* New York: McGraw-Hill. [LET]

Gulliksen, H. (1950) *Theory of mental tests.* New York: Wiley. [GMH, PEV]

Hall, W. S. (1978) Comments. In: *Achievement testing of disadvantaged and minority students for educational program evaluation*, ed. M. J. Wargo & D. R. Green. Monterey, Calif.: CTB/McGraw-Hill. [DRG]

Hall, W. S., Cole, M., Reder, S., & Dowley, G. (1977) Variations in young children's use of language: Some effects of setting and dialect. In: *Discourse production and comprehension*, ed. R. O. Freedle, Discourse processes: Advances in research and theory, vol. 1. Norwood, N.J: Ablex. [DRG]

Harrington, G. M. (1975) Intelligence tests may favour the majority groups in a population. *Nature* 258:708–709. [JH, ARJ, CRR]

Hayek, F. A. (1955) *The Counter-revolution of science.* London: Free Press. [WRH]

Hearnshaw, L. S. (1979) *Cyril Burt: Psychologist.* London: Hodder and Stoughton. [DDD]

Hebb, D. O. (1949) *The organization of Behavior.* New York: Wiley. [ARJ]

Henderson, N. B., Butler, B. V. & Goffeney, B. (1969) Effectiveness of the WISC and Bender Gestalt in predicting arithmetic and reading achievement for white and non-white children. *Journal of Clinical Psychology* 25:268–71. [ARJ]

Herrnstein, R. J. (1971) *IQ in the meritocracy.* Boston: Little, Brown & Co. [RBC]

Hirsch, J. (1963) Behavior genetics and individuality understood: Behaviorism's counterfactual dogma blinded the behavioral sciences to the significance of meiosis. *Science* 142:1436–1442. [JH]

(1967) Epilog: Behavior-genetic analysis. In: Behavior-*genetic analysis*, ed. J. Hirsch. New York: McGraw-Hill. [JH]

(1970) Behavior-genetic analysis and its biosocial consequences. *Seminars in Psychiatry* 2:89–105. [JH, DW]

(1976) Jensenism: The bankruptcy of "science" without scholarship. *United States Congressional Record*, 122: No. 73, E2671–2672: No. 74 E2693-2695: No. 75, E2703-2705, E2716-2718, E2721-2722 (originally published in *Educational Theory*, 1975, 25:3–27, 102). [JH]

Hirsch, J., McGuire, T. R., & Vetta, A. (1980) Concepts of behavior genetics and misapplications to humans. In: *The evolution of human social behavior*, ed. J. Lockard. New York: Elsevier-North Holland. [JH]

Hirsch, J., & Vetta, A. (1978) Gli errori concettuali dell'analisi genetico-comportamentale (The misconceptions of behavior genetics). *Richerche di psicologia*, Franco Angeli Editore. [JH]

Hirsch, N. D. M. (1930) An experimental study upon three hundred children over a six year period. *Genetic Psychology Monographs* 7, no. 6. [AV]

Hodos, W., & Campbell, C. B. G. (1969) *Scala naturae:* Why there is no theory in comparative psychology. *Psychological Review* 76:337–350. [JH, DW]

Horn, J. L., & Cattell, R. B. (1978) A check on the theory of fluid and crystallized intelligence, with description of new subtest designs. *Journal of Educational Measurement* 15:139–164. [RBC]

Horst, P. (1968) *Psychological measurement and prediction.* Belmond, Calif.: Brooks/Cole. [GMH]

Humphreys, L. G. (1973) Statistical definitions of test validity for minority groups. *Journal of Applied Psychology* 58:1–4. [RBC]

Hunt, J. McV. (1969) Has compensatory education failed? *Harvard Educational Review* 39:449–483.[AV]

(1979) Psychological development: Early experience. *Annual Review of Psychology* 30:103–143. [LET]

Hunter, J. E. (1975) A critical analysis of the use of item means and item-test correlations to determine the presence or absence of content bias in achievement test items. Paper presented to the National Institute of Education conference on Test Bias, Annapolis, Md. [CRR]

Ingle, D. J. (1978) Fallacies in agruments on human differences. In: *Human Variation: The biopsychology of age, race, and sex*, ed. R. T. Osborne, C. E. Noble, & N. Weyl. New York: Academic Press. [ARJ]

Irvine, S. H. (1969) The factor analysis of African abilities and attainments: Constructs across cultures. *Psychological Bulletin* 71:20–32. [PK]

Jencks, C. (1972) *Inequality: A reassessment of the effect of family and schooling in America*. New York: Harper and Row. [DLE]

(1979) *Who gets ahead? The determinants of economic success in America*. New York: Basic Books. [JE, ARJ]

Jensen, A. R. (1967) Estimations of the limits of heritability of traits by comparison of monozygotic and dizygotic twins. *Proceedings of the National Academy of Sciences* 50:149–156. [JH]

(1969) How much can we boost IQ and scholastic achievement? *Harvard Educational Review* 39:1–123. [AMC, JE, RAG, RTO, PEV, AV, DW]

(1969a) Reducing the heredity-environment uncertainty. *Harvard Educational Review* 39:449–483. [AV]

(1970) IQ's of identical twins reared apart. *Behavior Genetics* 1:133–148. [DDD]

(1971) Can we and should we study race differences? In: *Race and intelligence*, ed. C. L. Brace, G. R. Gamble, & J. T. Bond, Anthropological Studies No. 8, pp. 10–31. Washington, D.C.: American Anthropological Association. [CLB]

(1972) *Genetics and education*. New York: Harper & Row. [RBC, DDD, RTO, AV]

(1973) *Educability and group differences*. New York: Harper and Row. [RBC, DDD, JE, RAG, ARJ]

(1974a) How biased are culture-loaded tests? *Genetic Psychology Monographs* 90:185–244. [NB]

(1974b) Kinship correlations reported by Sir Cyril Burt. *Behavior Genetics* 4:1–28. [JE, ARJ]

(1978) The current status of the IQ controversy. *Australian Psychologist* 13:7–27. [RAG, ARJ]

(1979) The nature of intelligence and its relation to learning. *Journal of Research and Development in Education* 12:79–95. [ARJ]

(1980) *Bias in mental testing*. New York: Free Press. [cited by ARJ and all commentators]

(1980b) Chronometric analysis of mental ability. *Journal of Social and Biological Structures* [ARJ]

(1981) *Straight talk about mental tests*. New York: Free Press. [ARJ]

Jinks, J. L., & Fulker, D. W. (1970). Comparison of biometrical genetical, MAVA and classical approaches to the analysis of human behaviour. *Psychological Bulletin* 70:311–349. [AV]

Johnson, D. M. (1948) Applications of the standard-score IQ to social statistics. *Journal of Social Psychology* 27:217–227. [RAG]

Jordan, W. (1978) *White over black: American attitudes toward the Negro, 1550–1812*. Baltimore, Md.: Penguin. [DLE]

Judson, H. F. (1979) *The eighth day of creation*. New York: Simon & Schuster. [ARJ]

Kamin, L. J. (1974) *The science and politics of IQ*. New York: Halsted. [JE, DW]

Karlsson, J. L. (1978) *Inheritance of creative intelligence*. Chicago: Nelson-Hall. [RTO]

Kempthorne, O. (1977) Fundamentals of path analysis and population genetics. Review of C. C. Li (1975), *Path analysis: A primer*. Pacific Grove, Calif.: Boxwood Press. [JH]

(1978) Logical, epistemological and statistical aspects of Nature-Nurture data interpretation. *Biometrics* 34:1–23. [DDD, JH, AV]

Knapp, R. R. (1960) The effects of time limits on the intelligence test performance of Mexican and American subjects. *Journal of Educational Psychology* 51:14–20. [RBC]

Laboratory of Comparative Human Cognition. (in press) Intelligence as cultural practice. In: *Handbook of human intelligence*, ed. R. J. Sternberg. New York: Cambridge University Press. [RJS]

Lewontin, R. C. (1970) Race and intelligence. *Bulletin of the Atomic Scientists* March:2–8. [JE, DW]

(1974) The analysis of variance and the analysis of cause. *American Journal of Human Genetics* 26:400–411. [DW]

Li, C. C. (1955) *Population genetics*. Chicago: University of Chicago Press. [JH]

(1975) *Path analysis: A primer*. Pacific Grove, Calif.: Boxwood Press. [ARJ]

Linn, R. L. (1973) Fair test use in selection. *Review of Educational Research* 43 2:139–161. [ARJ]

Loehlin, J. C., Lindzey, G., & Spuhler, J. N. (1975) *Race differences in intelligence*. San Francisco: Freeman. [RBC]

Loehlin, J. C., & Vandenberg, S. G. (1968) Genetic and environmental components in the covariation of cognitive abilities: An additive model. In: *Progress in human behavior genetics*, ed. S. G. Vandenberg, pp. 261–285. Baltimore, Md.: Johns Hopkins University Press. [SGV]

Lord, F. M., & Novick, M. R. (1968) *Statistical theories of mental test scores*. Reading, Mass.: Addison-Wesley. [GMH]

Lynn, R. (1978) Ethnic and racial differences in intelligence: International comparisons. In: *Human variation: The bio-psychology of age, race and sex*, ed. R. T. Osborne, C. E. Noble, & N. Weyl. New York: Academic Press. [RBC]

McClelland, D. C. (1974) Testing for competence rather than for "intelligence." In: *The new assault on equality*, ed. A. Gartner, C. Greer, & F. Riessman. New York: Harper and Row. [DLE]

McDill, E. L., McDill, M. S., & Sprehe, J. T. (1969) *Strategies for success in compensatory education*. Baltimore, Md.: Johns Hopkins University Press. [RAG]

McGuire, T. R., & Hirsch, J. (1977) General intelligence (g) and heritability (H^2, h^2). In: *The structuring of experience*, ed. E. C. Uzgiris & G. Weizmann. New York: Plenum Press. [JH]

McGurk, F. C. J. (1967) The culture hypothesis and psychological tests. In: *Race and Modern Science*, ed. R. E. Kuttner. New York: Social Science Press. [ARJ]

McNemar, Q. (1942) *The revison of the Stanford-Binet scale*. Boston: Houghton Mifflin. [RAG]

(1949) *Psychological statistics*. New York: Wiley. [ARJ]

Maxwell, A. E. (1972) Factor analysis; Thomson's sampling theory recalled. *British Journal of Mathematical and Statistical Psychology* 25:1–21. [ARJ]

Medawar, P. B. (1977) Unnatural science. *New York Review of Books*, February:13–18. [DDD]

Merz, W. R. (1978) Test fairness and test bias: A review of procedures. In: *Achievement testing of disadvantaged and minority students for educational program evaluation*, ed. M. J. Wargo & D. R. Green. Monterey, Calif.: CTB/McGraw-Hill. [DRG]

Messé, L. A., Crano, W. D., Messé, S. R., & Rice, W. (1979) Evaluation of the predictive validity of tests of mental ability for classroom performance in elementary grades. *Journal of educational psychology* 71:233–241. [ARJ, LEL]

Mulaik, Stanley A. (1972) *The Foundations of factor analysis*. New York: McGraw-Hill. [OK]

Munday, L. (1965) Predicting college grades in predominantly Negro colleges. *Journal of Educational Measurement* 2:157–160. [HMB]

News and views. (1980) *Nature* 284:402. [JH]

Nichols, R. C. (1979) Policy implications of the IQ controversy. In: *Review of Research in Education*, ed. L. S. Shulman, Vol. 6. Itasea, Ill.: Peacock. [ARJ]

Nunally, J. C. (1967) *Psychometric theory*. New York: McGraw-Hill. [GMH]

Rozeboom, W. W. (1966) *Foundations of the theory of prediction*. Homewood, Ill.: Dorsey. [GMH]

Okun, A. M. (1976) Equal rights but unequal incomes. *New York Times Magazine*, July 4:101ff. [RAG]

Ozenne, D. G., Van Gelder, N. C., & Cohen, A. J. (1974) *Achievement test restandardization: Emergency School Aid Act national evaluation*. Santa Monica, Calif: System Development Corporation. [DRG]

Peckham, R. F. (1979) Opinion. Larry P. et al. v. Wilson Riles et al. United States District Court, Northern District of California, San Franscisco. No. C-71-2270RFP. [RAG, RTO, CRR]

Plomin, R., & DeFries, J. C. (1980) Genetics and intelligence: Recent data. *Intelligence* 4:15–24. [RAG, ARJ]

Pitt, C. C. V. An experimental study of the effects of teacher's knowledge or incorrect knowledge of pupil IQ's on teachers' attitudes and practices and pupils' attitudes and achievement. *Dissertation Abstracts* 16:2387–2388. [RR]

Reschley, D. J. & Reschley, J. E. (1979) Validity of WISC-R factor scores in predicting achievement and attention for four sociocultural groups. *Journal of School Psychology* 17:335–61. [ARJ]

Reynolds, C. R. (in press a) The problem of bias in psychological assessment. In: *The Handbook of School Psychology*, ed. C. R. Reynolds & T. B. Gutkin. New York: John Wiley and Sons. [CRR]

(in press b) Differential construct validity of intelligence as popularly measured: Correlations of raw scores with age on the WISC-R for blacks, whites, males, and females. *Intelligence: A Multidisciplinary Journal*. [CRR]

Rosenthal, R. (1978) How often are out numbers wrong? *American Psychologist* 33:1005–1008. [RR]

Rosenthal, R., & Jacobson, L. (1968) *Pygmalion in the classroom*. New York: Holt, Rinehart, & Winston. [ARJ, RR]

Rosenthal, R., & Rubin, D. B. (1971) Pygmalion reaffirmed. In: *Pygmalion reconsidered*, ed. J. D. Elashoff & R. E. Snow. pp. 139–155. Worthington, Ohio: Charles A. Jones. [RR]

(1978) Interpersonal expectancy effects: The first 345 studies. *Behavioral and Brain Sciences* 3:377–386. [RR]

(1979) A note on percent variance explained as a measure of the importance of effects. *Journal of Applied Social Psychology* 9:395–396. [RR]

(1980) A simple, general purpose display of magnitude of experimental effect. Unpublished manuscript, Harvard University. [RR]

Rudner, L. M. (1977) Efforts toward the development of unbiased selection and assessment instruments. Paper presented at the Third International Symposium on Educational Testing, University of Leyden, The Netherlands. [DRG]

Samuel, W. (1977) Observed IQ as a function of test atmosphere, tester expectation, and race of tester: A replication for female subjects. *Journal of Educational Psychology* 69:593–604. [ARJ, RR]

Scarr, S., & Carter-Saltzman, L. (in press) Genetics and intelligence. In: *Handbook of human intelligence*, ed. R. J. Sternberg. New York: Cambridge University Press. [RJS]

Scarr-Salapatek, S. (1971) Race, social class and IQ. *Science* 174:1285–1295. [DW]

Savage, I. R. (1975) Review of J. C. Loehlin, G. Lindzey, & J. N. Spuhler, 1975, *Race differences in intelligence. Proceedings of the National Academy of Education* 2:1–37. [JH]

Seaver, W. B. (1973) Effects of naturally induced teacher expectancies. *Journal of Personality and Social Psychology* 28:333–342. [RR]

Sewall, G., & Howard, L. (1979) A high grade for Head Start. *Newsweek* Oct. 8:102 p. 102. [JH]

Shuey, A. M. (1966) *The testing of Negro intelligence*. 2nd ed. New York: Social Science Press. [RBC, RAG]

Silverman, B. I., Barton, F., & Lyon, M. (1976) Minority group status and bias in college admission criteria. *Educational and Psychological Measurement* 36(2):401–407. [HMB]

Spearman, C. (1904) "General intelligence": Objectively determined and measured. *American Journal of Psychology* 15:201–292. [ARJ]

(1914) The heredity of abilities. *Eugenics Review* 6:219–237. [JH]

Spencer, H. (1899) *The principles of psychology*, I. 3rd ed. New York: D. Appleton. [DLE]

Stafford, R. E. (1961) Sex differences in spatial visualization as evidence of sex-linked inheritance. *Perceptual and Motor Skills* 13:300–308. [RBC]

Terman, L. M. (1916) *The measurement of intelligence*. Boston: Houghton Mifflin. [DLE]

(1917) Feeble-minded children in the public schools of California. *School and Society*, 5:161–165. [DLE]

Terman, L. M., & Merrill, M. A. (1960) *Stanford-Binet intelligence scale: Manual for the third revision, Form L-M*. Boston: Houghton Mifflin. [RAG]

Thoday, J. M. (1973) Review of A. R. Jensen, 1973, *Educability and group differences*. *Nature* 245:418–420. [AV]

Thomas, G., Alexander, K., & Eckland, B. K. (1979) Access to higher education: The importance of race, sex, social class, and academic credentials, *School Review* 87:133–56. [BKE]

Thorndike, E. L. (1905) Measurements of twins. In: *Archives of philosophy, psychology and scientific methods*, ed. J. M. Cattell & F. J. E. Woodbridge. New York: Science Press. Sept. 1905–July 1906, 1:1–63. [RTO]

Tillyard, E. M. W. (1944) *The Elizabethan world picture*. New York: Random House. [DLE]

Triandis, H., Malpass, R. S., & Davidson, A. R. (1971) Cross-cultural psychology. In: *Biennial review of anthropology*, ed. B. J. Siegel. Stanford: Stanford University Press. [PK]

Tuddenham, R. D. (1948) Soldier intelligence in World Wars I and II. *American Psychologist* 3:54–56. [RAG]

Urbach, P. (1974) Progress and degeneration in the IQ debate. *British Journal of the Philosophy of Science* 25:99–135, 235–59. [ARJ, RTO]

Vandenberg, S. G. (1965) Multivariate analysis of twin differences. In: *Methods and goals in human behavior genetics*, ed. S. G. Vandenberg. pp. 29–40. New York: Academic Press. [SGV]

(1968) The nature and nurture of intelligence. In: *Genetics*, ed. D. C. Glass, pp. 3–58. New York: Rockefeller University Press & Russell Sage Foundation. [SGV]

(1973) Comparative studies of multiple factor ability measures. In: *Theoretical problems in multivariate research*, ed. J. Royce. New York: Academic Press. [SGV]

Van Valen, L. (1974) Brain size and intelligence in man. *American Journal of Physical Anthropology* 40:417–423. [ARJ]

Vernon, P. E. (1979) *Intelligence: Heredity and environment*. San Francisco: W. H. Freeman. [ARJ]

Vetta, A. (1975) Regression to the mean. *Social Biology* 22:87–88. [AV]

(1977) Genetical concepts and IQ. *Social Biology* 24:166–169. [AV]

(1977a) Estimation of heritability from IQ data on twins. *Nature* 266:279. [JH]

(1980) Concepts and issues in the IQ debate. *Bulletin of the British Psychological Society*, in press. [AV]

Vygotsky, L. S. (1978) In: *The development of higher psychological processes*, ed. M. Cole, V. John-Steiner, S. Scribner, & E. Souberman. Cambridge, Mass.: Harvard University Press. [RJS]

Wahlsten, D. (1978) Behavioral genetics and animal learning. In: *Psychopharmacology of aversively motivated behavior*, ed. H. Anisman & G. Bignami. New York: Plenum.

(1979) A critique of the concepts of heritability and heredity in behavioral genetics. In: *Theoretical advances in behavior genetics*, ed. J. R. Royce & L. P. Mos. Germantown, Md.: Sithoff & Noordhoff. [DW]

Warburton, F. J. (1951) The ability of the Ghurkha recruit. *British Journal of Psychology*, 42:123–133. [PK]

Wallach, M. A. (1976) Tests tell us little about talent. *American Scientist* 64:57–63. [DLE]

Weiss, R. H. (1972) Möglichkeiten und Grenzen des "Culture Fair Intelligence Tests" (CFT) in der Schulbahnberatung. In: *Festschrift for Dr. D. W. Arnold*. Frankfurt: Peter Lang. [RBC]

Weizmann, F. (1970) Correlational statistics and the nature-nurture problem. *Science* 171:589. [JH]

Willerman, L. (1979) Effects of families on intellectual development. *American Psychologist* 34:923–929. [ARJ]

Williams, T. R. (ed.) (1975) *Psychological anthropology*. The Hague: Mouton. [PK]

Wilson, J. R., & Vandenberg, S. G. (1978) Sex differences in cognition: Evidence from the Hawaii Family Study. In: *Sex and behavior: Status and prospectus*, ed. T. E. McGill, D. A. Dewsbury, & B. D. Sachs, pp. 317–336. [SGV]

Wilson, R. S. (1978) Synchronies in mental development: An epigenetic perspective. *Science* 202:939–948. [DW]

Wober, M. (1973) Towards an understanding of the Kigmda concept of intelligence. In: *Culture and cognition: Readings in cross-cultured psychology*, ed. J. W. Berry & P. R. Dasen. London: Methuen. [PK]

Wright, S. (1921) Correlation and scusation. *Journal of Agricultural Research* 20:557–585. [ARJ]

Yerkes, R. M. Psychological examining in the United States Army. *Memoirs of the National Academy of Sciences* 15:1–890. [RAG]

Batshaw, L. M., Roan, Y., Jung, A. L., Rosenberg, L. A. & Brusilow, S. W. (1980) Cerebral dysfunction in asymptomatic carriers of ornithine transcarbamylase deficiency. *New England Journal of Medicine* 302, Vol. 8: 482–485. [FV]

Thalhammer, O., Pollak, A., Lubec, G. & Königshofer, H. (1980) Intracellular Concentrations of Phenylalanine, Tyrosine and alpha-Aminobutyric Acid in 13 Homozygotes and 19 Heterozygotes for Phenylketonuria (PKU) Compared with 26 Normals. *Human Genetics* 54:213–216. [FV]

Thalhammer, O., Lubec, G. & Königshofer, H. (1979) Intrevellular phenylalanine and tyrosine concentrations in 19 heterozygotes for phenylketonuria (PKU) and 26 normals. Do the higher values in heterozygotes explain their lowered intellectual level? *Human Genetics* 49:333. [FV]

Thalhammer, O., Havelec, L., Knoll, E. & Wehle, E. (1977) Intellectual level (IQ) in heterozygotes for phenylketonuria (PKU). Is the PKU gene also acting by means other than phenylalanine blood level elevation? *Human Genetics* 38:285. [FV]

Vogel, F. & Motulsky, A. G. (1979) *Human Genetics—Problems and Approaches*. New York: Springer-Verlag. [FV]