Siegel, S., Hinson, R. E. & Krank, M. D. (1978) The role of pre-drug signals in morphine analgesic tolerance: Support for a Pavlovian conditioning model of tolerance. *Journal of Experimental Psychology: Animal Behavior Processes* 4:188-96. [MSF]

Solomon, R. L. & Corbit, J. D. (1974) An opponent-process theory of motivation: I. Temporal dynamics of affect. *Psychological Review* 81:119-45. [MSF, NJM]

Ursin, H. (1980) Affective and instrumental aspects of fear and aggression. In: *Functional states of the brain: Their determinants*, ed. M. Koukkou, D. Lehmann, & J. Angst. Amsterdam: Elsevier. [HU]

Ursin, H., Wester, K. & Ursin, R. (1967) Habituation to electrical stimulation of the brain in unanesthetized cats. *Electroencephalography and Clinical Neurophysiology* 23:41-49. [HU]

Zener, K. (1937) The significance of behavior accompanying conditioned salivary secretion for theories of the conditioned response. *American Journal of Psychology* 50:384-403. [MSF]

On Arthur R. Jensen (1980) Précis of *Bias in Mental Testing.* BBS 3:325-371.

Abstract of the original article: 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.

The validity of Jensen's statistical methods

Richard B. Darlington and Carolyn M. Boyce
Department of Psychology, Cornell University, Ithaca, N.Y. 14853

To assess the adequacy of Arthur Jensen's (1980a) statistical methods, we selected for close examination one small section of *Bias in Mental Testing* - the notes at the end of Chapter 9. These notes seem to form the psychometric and statistical basis for much of the book's material on racial differences.

We further restricted this commentary to the 13 of the 17 notes we defined as "elementary," meaning that the note's major point was explicitly or apparently taken from a textbook at the undergraduate or introductory graduate level, or could be proven false by reference to such a book. By this criterion Notes 3, 12, 14, and 15 were not elementary.

The content of the 13 elementary notes is such that one might expect one or two errors at the very most. Yet we found nine major errors and one minor one.

Jensen's Note 1 is a discussion of the sampling error of simple regression lines. Note 1 includes formulas for the standard error of a regression slope b, the standard error of the Y-intercept a, and a test for the significance of the difference between two independent values of a. All three of these formulas are incorrect. This is especially noteworthy because the most widely used single definition of test bias involves testing whether regression lines in two cultural groups have the same value of a. One of the errors in Note 1 concerned that test.

Jensen's Note 1 gives the standard error of a regression weight b as $(S_Y/S_X)\sqrt{(1 - r^2_{XY})/(N - 1)}$. The correct form, given in numerous texts, contains $(N - 2)$ where Jensen has $(N - 1)$.

The standard error that Jensen gives for the Y-intercept of a regression line (the value of $\hat{Y}$ at $X = 0$) is in fact the standard error of $\hat{Y}$ at $X = \bar{X}$. The former standard error can be many times the latter.

The third error in Note 1 concerns a method for testing the significance of the difference between the Y-intercepts of regression lines in two independent groups of people, and states that this is algebraically equivalent to analysis of covariance. Even casual inspection reveals the nonequivalence. In Jensen's method the final value of t is affected by which of the two groups is labeled Group A and which Group B. This is clearly undesirable, since such labeling may be purely arbitrary. Analysis of covariance lacks this undesirable property. Thus Jensen's method is not only nonequivalent to analysis of covariance, but is clearly inferior as a test of the equality of two intercepts. We easily constructed examples in which the difference between races is either nonsignificant at the .05 level, or significant beyond the .01 level, depending on which racial group is called group A. Our examples contained the most standard assumptions: equal variances, correlations, slopes, and standard errors of estimate between groups; normal distributions; and a difference of one standard deviation on X between races - the value Jensen says is typically observed. All that is required to produce anomalous results is to assume the two sample sizes differ substantially - a circumstance that occurs frequently when comparing two racial groups.

Note 4 presents a formula from McNemar (1949) for correcting a correlation coefficient for range restriction. It mentions that McNemar warns against a double correction - correcting first for restriction on X and then correcting this value for restriction on Y. But the note explicitly recommends ignoring McNemar's warning. That is simply incorrect; the formula was derived to make only one correction necessary or correct.

Note 6 is a discussion of internal consistency reliability, with emphasis on the Kuder-Richardson 20 formula. It says, "The significance of the difference between the [K−R 20] reliability coefficients is determined by the same method used in testing the difference between two correlation coefficients, using Fisher's z transformation of r (see Guilford, 1956, pp. 182 - 183, 194)." The pages cited in Guilford contain a thoroughly ordinary discussion of the Fisher z, with no mention or even hint that it can be applied to K−R 20 values.

We believe that the exact sampling distribution of a K−R 20 reliability is unknown, but we have attempted various approximations. Jensen's formula tentatively appears to be fairly accurate if k (the number of items in a test) is large. However, if k is small, then contrary to Jensen's suggestion, the formula is highly inaccurate. Consider the case in which $k = 2$ and the two items in the test have equal standard deviations and correlate .10 with each other. Then the K−R 20 reliability is .1818. Jensen's statement implies that we could use the Fisher z to test the hypothesis of zero internal consistency by testing the K−R 20 value against 0. If $N = 200$ then this test yields $p = .005$ one-tailed. But the hypothesis of zero internal consistency is really the hypothesis that the two test items correlate zero, and thus should be tested with $r = .10$, not $r = .1818$, since .10 is the observed correlation between the two test items. This test yields $p = .08$ one-tailed by either the t test or the Fisher z. Thus the correct p is .08 while Jensen's suggested test gives $p = .005$.

Note 7 considers the problem of testing whether the items in a test of k items are more highly intercorrelated in one cultural group than in another. The particular measure of correlation Jensen discusses is phi/phi$_{\max}$, but we are not concerned here with the particular measure of correlation used, so we shall simply call it r. Note 7 suggests taking the $k(k - 1)/2$ values of r computed in each cultural group, and comparing the two sets of r values with a matched-pairs t test. Since n in this test would be the number of correlations, the number of subjects used in computing the correlations does not enter into the test at all. Thus this test does not at all assess whether the differences between cultural groups could be due to sampling error caused by small samples of subjects. The n in the test is the number of correlations. But these $k(k - 1)/2$ correlations are in no sense mutually independent. First, the items themselves are intercorrelated. Second, even if the items were uncorrelated, the correlations themselves would be interdependent because $k - 1$ correlations are computed for each item. Note 7 illustrates a problem that reappears throughout the book: incorrect choice of sampling unit for hypothesis tests.

Note 9 incorrectly deleted the brackets in the formula for the Fisher r-to-z transformation. The correct form is $z = (½)[ln(1 + r) - ln(1 - r)]$. The same error appears in Guilford (1956), which Jensen quotes frequently. There is no way to know whether Jensen has ever actually used the formula in its incorrect form. But if he has, the error is serious; the values of z calculated by the correct and incorrect forms are very different.

Note 10 refers to a discussion on page 438 which reflects the same sort of confusion as Note 7: little or no sense of the requirement of independence in applying hypothesis tests.

Note 13 presents a grossly incorrect method for testing the difference between two proportions observed in independent samples A and B. Jensen's method is based on transforming each p to a z by the probit transformation–that is, finding the z in a normal table that corresponds to the given value of p. The test in Note 13 is based on Jensen's belief that "the standard error of z is always $SE_z = 1/\sqrt{N} - 1$" (1980a, p. 440). The test suggested in Note 13 is then equivalent to

$$z = (z_A - z_B)/\sqrt{1/(N_A - 1) + 1/(N_B - 1)}.$$

This would be a reasonable test if the quote were correct. But it is not; the standard error of a z-transformation refers to the variation of z *across* samples, and has nothing to do with the fact that ordinary z scores have unit variances *within* samples. The easiest way to show the seriousness of this error is to compare the test in Note 13 to a standard test for the differences between two proportions. If two proportions are .05 and .10 in independent samples of 100 people each, then the standard Fisher 2×2 test gives $p = .14$ one-tailed, while Jensen's test gives $p = .005$.

The final note, Note 17, discusses the well-known phi measure of correlation between two dichotomous items, and the less well-known correlation measure Q. If the four cell frequencies are A, B, C, D, then

$$\text{phi} = (AD - BC)/\sqrt{(A + B)(C + D)(A + C)(B + D)}$$

and

$$Q = (AD - BC)/(AD + BC).$$

Note 17 states: "Q is monotonically related to the phi coefficient," meaning that if phi is higher in sample 1 than in sample 2 then Q will also be higher or at least as high. But this is grossly incorrect. For instance, let A, B, C, D respectively be 50, 0, 900, 50 in sample 1, and 300, 200, 200, 300 in sample 2. Then phi is .053 in sample 1 and .200 in sample 2, making the association appear much higher in sample 2. But Q is 1.00 in sample 1 and .385 in sample 2, making the association appear much higher in sample 1. Actually phi and Q measure two different aspects of association. For right-wrong items, phi measures the degree to which two items assess the same ability at the same difficulty level, while Q measures the degree to which the two items assess the same ability regardless of difficulty levels.

We have discussed 10 noteworthy psychometric and statistical errors in 13 elementary notes covering less than five pages of Jensen's latest book. Nine of the 10 errors could lead to grossly incorrect results in data analysis; the tenth involved using $(N - 1)$ where $(N - 2)$ is correct. All 10 errors can be detected merely by reference to undergraduate and introductory graduate level textbooks. The errors are basic, and appear to be central to Jensen's major arguments. This analysis alone would seem to provide substantial ground for doubting Jensen's major conclusions, apart from any further critical considerations.

The challenge is unmet

Jerry Hirsch[a] and Timothy P. Tully[b]
[a]*Departments of Psychology and of Ecology, Ethology, and Evolution, and Institutional Racism Program, University of Illinois, Urbana-Champaign, Ill. 61820 and* [b]*Department of Biology, Princeton University, Princeton, N.J. 08544*

In his Response to the BBS multiple review of *Bias in Mental Testing* (Jensen 1980a), Jensen (1980b, p. 360) says that "of the total of 27 reviewers [actually there were 28 reviews, 32 reviewers], 18 express agreement with the book's main conclusions. The remaining nine are either noncommittal or address side issues, but not one directly challenges the main conclusions." We have, in fact, challenged the main conclusion of Jensen's book by describing the experimental design of Harrington (1975), and we have addressed a critical side issue, genetics, through our experiences in that field. Because Jensen's remarks about our review contain serious errors, we have prepared the following clarifications for Continuing Commentary.

Jensen has misconstrued our discussion of Harrington's (1975) study. When Jensen speaks of "generalization to humans," he misses the main point of our quotation from Harrington, who had explicitly stated that his "experiment was an empirical test of common psychometric assumptions and procedures. Generalization is therefore to those assumptions and procedures." In other words, this is methodological generality, in the same sense that the methods of statistical analysis apply equally well to plant, animal, and human data. By using the methods of human test construction and standardization and by manipulating in a balanced systematic *experiment* the racial composition of the standardizing population(s) essential for the construction of test(s), Harrington has demonstrated the fundamental influence of that variable (racial composition) on the results obtained by administering tests so constructed to different races, that is, that "genotype-environment interactions . . . do in fact affect item selection and that these effects in turn contribute to the phenomena of group differences" (Harrington 1975, p. 708). Each of the six races had different sets of items on which they performed well. The results were interpreted to mean that when such a condition exists, the "majority will score higher than minorities as a general artifact of test-construction procedures" (Harrington 1975, p. 709; Hirsch 1976, p. 8, showed "that... since none of the populations contained a majority, the largest group in each population was a 'plurality' or 'principal group'"). He reported only moderately strong, significant but *not perfect*, correlations between group performance and level of representation in the standardizing population. Such a correlation, however, can occur even if a minority group scores slightly higher than the majority (e.g. Japanese score higher than whites). Furthermore, the possibility exists for a minority to outscore a majority on the same set of items. In that case, standard test construction procedures (i.e. item selection) would yield a test for the standardizing population on which a minority would outperform a majority group. In short, the implications of Harrington's study do not pose a contradiction for human IQ testing. In fact, his study may explain why, as Jensen (1980b, p. 361) says, "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."

Harrington has produced a design that directly manipulates the composition of the standardization population in an experimental framework. His design demonstrates that so-called bias is intrinsic and truly unavoidable in the methods of test construction. *Verbal* protestations about a "standardization fallacy" (Longstreth 1980, p. 350) should now yield to experimental evidence.

Another reviewer, Reynolds (1980, p. 352), has recognized and remarked about the importance of Harrington's (1975) study: "If... correct, then... 100 years... of psychological research in human differences... must be dismissed as confounded, contaminated, or otherwise artifactual." Also, he has since arranged for publication of a more complete exposition of this fundamental analysis (Harrington 1982). Unfortunately, Reynolds's comment (1980, p. 352) that according to "the cultural test bias hypothesis... all... group score differences are an artifact of current psychometric methodology," with its accompanying reference to Harrington (p. 352) is too easily interpreted to imply that all critics of testing believe that (1) *only* differences in culture might produce test-score differences among individuals or groups, (2) it is possible to create an unbiased test, and (3) in the absence of cultural differences, such an unbiased test might show all individuals and groups obtaining equal scores.

The foregoing does *not* represent our interpretation of Harrington's work, which is only pertinent to "cultural bias" by analogy. Individual and group differences are not created by tests, but since test construction is based on the reactions of individuals and groups who will differ for either genetic or environmental reasons or both, the very idea that any test so constructed might be neutral or unbiased is meaningless. That is, tests are constructed according to criteria for selecting items based upon the responses of individuals, all of whom differ. As Harrington has so effectively shown, psychometric procedures will select from an available pool different combinations of items depending upon which individuals are included in the standardizing population. This in turn will influence their relative standing when different populations are compared with respect to their performance on a given test, which is similar to the point being made in his title and penultimate paragraph by Dorfman (1980).

Explanation of the fundamentals involved with respect to both the cultural and the genetic influences have been presented in *Science* by Bohannan and by Hirsch:

> There is no possibility of any "intelligence" test *not* being culturally biased. The content of an intelligence test must have something to do with the ideas or the muscle habits or with habitual modes of perception and action of the people who take the test. All these things are culturally mediated or influenced in human beings (even man's actions as a mammal or a vertebrate are given cultural evaluations that influence the behavior itself). This is not a dictum or a definition - it is a recognition of the way in which cultural experience permeates everything human beings perceive and do. (Bohannan 1973, p. 115)

The foregoing in no way precludes the effect of a simultaneous bias attributable to the ubiquitous genetic diversity, which also influences all human behavior (J. Hirsch 1963). That the complexity thus involved defies description by simple quantitative models makes its reality no less scientific.

We are surprised that the psychometric community has been so reluctant to acknowledge the implications of Harrington's experimental design and to use it with human IQ tests, where experimentation is so lacking. Do they really believe, as William James once opined, that "empiricism is an awful waste of time"?

Jensen (1980b, p. 367) also speaks of "Harrington's finding, based on genetically different strains of rats... [as] fascinating." We must emphasize, as Harrington does, that the strains could differ environmentally and still produce analogous results. The fact that the strains differ, whether for genetic or environmental reasons, is the only precondition.

Kempthorne and Wolins (1980, p. 348) have stated: "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." Although we can agree conceptually with the latter half of their sentence, the former part is a conclusion based on confusion.

The paragraph in Kempthorne and Wolins (1980) that precedes the above quotation displays the logic employed. "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" (p. 348). We can label this comment the "predictive validity fallacy." There is an implicit assumption in much of the psychometric literature that the criterion variable is not biased. But if in fact a criterion variable does reflect the biased treatment of groups in society, then the predictor variable (i.e. the test) that best predicts the criterion score must also be biased. If one is willing to assume that the criterion variable is biased, then it follows directly that predictor tests that do not distinguish blacks from whites will not be valid predictors of "societally relevant criteria" for the blacks or whites. It is a fact that blacks have been discriminated against on many such relevant criteria (such as education, housing, employment, recreation, etc.). The inability of psychometricians to develop tests that are unbiased "despite arduous efforts" may stem from the complete reluctance to manipulate experimentally the properties of test construction procedures.

Harrington (1975) may constitute the first and *only* experiment in psychometrics to demonstrate conclusively a property of test standardization that can and does result in biased tests. The experimental design satisfies the assumptions and conditions discussed by Kempthorne (1978) that are necessary to demonstrate causation. *No* observational data in the IQ literature can stand up to the rigor or the conclusions of Harrington's experiment.

Our critique of Jensen (1980a) is from a behavior-genetic perspective. Central to this view is the importance, even the necessity, of breeding analysis to test any genetic hypothesis. It is from this perspective that Jensen's formula 6.10 remains an inappropriate application of a population parameter to the individual. Jensen (1980b, p. 361) states: "In quantitative genetics, the broad heritability h^2 can be conceived as the square of the Pearson correlation between genotypic and phenotypic values.... It follows that the regression of genotypic values on phenotypic values is... h^2." That interpretation holds only when there is no genotype-environment covariance. And, as Jensen also says: "It is precisely because the genotype becomes causally correlated with the environment and... 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" (1980b, p. 361).

As stated in Jensen's reference (Falconer 1960, pp. 132 - 33),

> this [phenotype-environment correlation] introduces a correlation between phenotypic value and environmental deviation; and, since genotype and phenotypic values are correlated, there is also a correlation between genotypic value and environmental deviation.... Thus... on practical grounds it is unavoidable, to regard any covariance that may arise from genotype - environment correlation as being part of the genotypic variance.

Jensen's heritability estimate exaggerates the genotypic variance by an unknown amount since genotypic values cannot be determined. One cannot have it both ways. Furthermore, the genetic model (Kempthorne 1957, Chapter 13) that Jensen states his formula 6.10 is based on does not apply to observational data. The model applies to an experimental design in which genetically different groups, that is, varieties, are randomized across environments. As Falconer (1960, p. 132) points out, "correlation between genotype and environment ... can usually be neglected in experimental populations, where randomization of environment is one of the chief objects of the experimental design." Jensen has incorrectly generalized to IQ data the genetic model that his formula 6.10 summarizes.

Furthermore, Jensen (1980b, p. 362) misunderstands "*Statistical interaction* of genotypic values and environmental values," and his conceptual demonstration there is inadequate. It represents another inappropriate application of an experimental design. To measure interaction it is necessary to replicate an array of genotypes, place them in an array of environments, and measure their distributions of responses (on some trait) in each environment; that is, each genotype is exposed to every environment, and all genotypes are exposed together to each environment. In this way each environment is replicated by its application to every genotype, and each genotype is replicated and exposed to every environment.

Jensen's own cited source (Falconer 1960, p. 134), states

> This interaction variance can be isolated and measured only under rather artificial circumstances. We may replicate genotypes by the use of inbred lines or F_1's, and replicate specific environments by the control of such factors as nutrition or temperature. Then an analysis of variance... will yield estimates of the genotypic variance (between genotypes), the environmental variance (between environments) and the variance attributable to interaction of genotypes with environments. The specific environments in such an experiment are, however, more in the nature of "treatments" because a population under genetical study would not normally encounter so wide a range of environments as that provided by the different treatments. It is therefore the genotype-environment interaction occurring within one such treatment that is relevant to the genetical study of a population, and this cannot be measured because the separate elements of the environment cannot be isolated and controlled.

Jensen's (1980b, p. 362) claim that "no one has yet been able to detect any significant component of IQ variance... associated with $G \times E$ interaction" is based on the error in the Jinks and Fulker (1970) source he cites - an error that, as Vetta (1980a, p. 357) pointed out, remains uncorrected in the literature. We are pleased to report that at long last Vetta's analysis and correction of Jinks and Fulker has finally been published and is now available in J. Hirsch (1981).

Once again, in his remarks about "the often repeated cliché," Jensen (1980b, p. 361) has shown confusion about heritability in general and about the relation of within-group heritability to differences between groups. The case he considers "of complete heritability ($h^2 = 1$) *within* each of two groups" absolutely precludes the condition next discussed of "environmental... factors showing variation *within* the groups," for the illustration of which it was adduced. The meaning of "complete heritability ($h^2 = 1$)" is the *absence* of relevant within-group environmental variation. However, nothing prevents two groups from having complete heritability under different environmental conditions, such as proper nutrition and adequate schools versus poor nutrition and inadequate schools, which in turn *might* affect a difference in group means. Nor does this preclude genetic differences affecting group means. Within-group heritability provides no information about between-group differences.

The discussions of the heritability of IQ in Jensen's (1980a) text and in his Response (1980b) seem mutually contradictory, and this remains unacknowledged and unreconciled. In his reply to us (1980b, p. 363), he opts for greater than zero heritability. In his reply to Vetta (1980a) he claims that "no one argues that the heritability... is of any particular value in general" (Jensen 1980b, p. 362) But, as both Wahlsten (1980, p. 359) and we (Hirsch, Beeman & Tully 1980, p. 346) have pointed out, Jensen had committed himself and "most geneticists" to agreeing that "80 percent or more of the IQ variance... is genetic." However, in his discussion of "genetics and heritability" (1980b, p. 361), Jensen also says, "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." If so, why does he continue advocating his belief by reconsidering old, observational data that can contribute nothing to an answer?

Neither Spearman (1914, pp. 220-21), McGuire and Hirsch (1977, p. 63), Hirsch, Beeman, and Tully (1980), nor J. McV. Hunt (1981) - all of whom discuss the same Spearman eugenic rationale for retaining a unitary *g* despite its genetically counterfactual status - has in any way considered or even suggested in his writings that "political views" were involved. Jensen's suggestion that the fruitfulness of *g* was being "decided on... political grounds" and therefore ad hominem (1980b, p. 363) is itself an ad hominem criticism (on this issue see Item Number 4 in Gillie 1980, p. 12, as well as his documentation of "falsehoods"; these appear in J. Hirsch 1981, p. 31 as well). Readers of this journal can consult these references and decide for themselves.

Our criticism did not consider the fruitfulness of *g*. It clarified what happens genetically, because Jensen has repeatedly propounded the knowledge claim, for which we feel there is no creditable evidence, that there is "a general factor in human intelligence, which we know to have a large genetic component" (1969, p. 456). He has stated that "The substantial heritability of... g-loaded tests is... proof of a biological basis for individual differences in *g*," (1980a, p. 251), and has remarked about "the extent that *g* is estimated by IQ" (1980b, p. 363), claiming (incorrectly) that a genetic component has already been demonstrated.

Jensen has misunderstood the issue involved in the alleged "biological unity" of *g*, which is neither more nor less of a unitary entity than is the economist's Gross National Product (GNP). The confusion of levels of analysis seems to be not ours but Jensen's, who, in our view, introduced genetics incorrectly into his discussion of *g*. In fact, in our review we suggested a genetic means - "the absence of random mating" - by which "traits with independent genetic correlates can show fortuitous correlations indefinitely, like the evidence for *g*," which has previously appeared in Hirsch (1967a, p. 125; 1967b, p. 433) but which Jensen has not yet taken into consideration.

We believe that heredity is important throughout human life, especially in intellectual functioning (J. Hirsch 1967c), just as before the event many of us *believed* that man would some day leave the earth and fly into space. But, belief is *not* science, even when called "hypothesis" or "theory." Nor can observational data prove causation.

Testing reveals a big social problem

Oscar Kempthorne[a] and Leroy Wolins[b]
[a]*Department of Statistics, Iowa State University, Ames, Iowa 50011 and*
[b]*Departments of Psychology and Statistics, Iowa State University, Ames, Iowa 50011*

Introduction. Arthur Jensen (1980a) has given us a rather large volume in which he addresses the question of whether the mental tests in common use are biased so that individual members of minority groups are treated unfairly.

The work has received a BBS multiple book review [*BBS* 3(3) 1980] as well as considerable discussion in the semi popular "literary" press. We (Kempthorne & Wolins 1980) contributed an abridged review to the first round in BBS because our original draft was much too long for copublication. However, our long review was regarded as having some utility (Jensen 1980b, p. 367), and there has been interest in seeing it published.

In the ensuing interval, we have read the other BBS reviews and those of serious writers in other publications. The paper that follows contains our detailed assessment. Some of our basic comments were already made in the first round in BBS. The present paper will include those that we regard as critical and necessary for an overall evaluation.

We consider it essential to look at the whole problem. Jensen's book contains his views on the origin of the questions about mental testing (which, we imagine, "triggered" his writing the book), on the theory of mental testing, and on the various types of tests that are used, as well as his examination of the potential existence of bias and his conclusion that the tests currently used do not have bias, as he defines it; finally, there are some general views on implications as Jensen perceives them.

Mental testing is under fire. There can be no question but that mental testing is indeed under fire. In Chapters 1 and 2, Jensen gives a very long description of his perception of the situation. Is it under fire because it has been used for tracking? It is obvious that tracking pervades education. In a community like Ames, Iowa, in which there are practically no racial minorities, such as blacks, Indians, Hispanics, or Orientals, there is tracking in the junior and senior high school. Some students study physics, analytic geometry, and trigonometry in 11th grade, and calculus and honors English in the 12th. Is this a result of tracking? Obviously, students who take these so-called high-level courses do so because for one reason or another they (or their guides) desire this. And also in general they have the mental abilities to make at least a reasonable stab at the material in those subjects at the current levels. In our colleges and universities, we are testing our students all the time; homework, mid-terms, and final exams. No one in his right mind questions the appropriateness of such testing, though there can be considerable differences in opinions on how much and what sort of testing should be done. And, of course, this leads to tracking.

It is obvious also that some sort of testing has always pervaded basic education. Every teacher presents ideas and then tests the pupils, if only to form an idea of whether or not the teaching has been successful.

Jensen tells us that opponents of standardized tests appear to take many different views such as: (a) testing is inhumane; (b) testing labels individuals in invidious ways; (c) testing is unfair; (d) testing is erroneous; (e) testing as practised tests only of narrow spectrum of mental abilities, and this favors some individuals unfairly insofar as test results are used to label and then to assign into groups, individuals who are strong or weak with respect to abilities the tests measure. One can go on and on. Jensen discusses the criticisms at length.

There can be no doubt that ability with respect to the three Rs, reading, writing, and arithmetic, can be tested. This is so obvious that a person who denies it must be regarded as irrational. But, of course, there must be rational realization of what is being done. We, the authors, could pass none of the tests if they were conducted in Hindi and the responses were to be given in Hindi. This remark brings out the whole matter of language and culture, and this is what much of the argument is about. If the tests used, say, in California, resulted in 50% of Caucasians scoring below 100, 25% of Caucasians below 90, and 50% of Hispanics below 100, 25% of Hispanics below 90, and so on, we would not have all the argument, legal, political, and otherwise, that we see pervading the whole area of education.

This does not happen, as we all know. If *any* of the common tests of mental ability or academic achievement is used, it is found that the proportion of individuals of any particular age who have a test score less than, say, 80 (or whatever), varies with ethnic origin. So tracking in Washington, D.C., schools resulted in a high degree of racial segregation. Jensen tells us (1980a, p. 27) that "the plaintiffs contended that the tracking system was discriminatory along racial and socioeconomic lines rather than in terms of capacity to learn." Then we are told that Judge J. Skelly Wright outlawed the tracking system as "'irrational and thus unconstitutionally discriminatory.'" Everyone who thinks about this situation and the mass of rhetoric associated with it will have his own reactions. Jensen appears to be fair in his description of the argument. We believe he makes a decent and fair account of various court cases on the whole problem.

Quite aside from Jensen's writings, we are exposed in the daily press and in our magazines to a sequence of problems, court cases, and judgments over most of the country. It is interesting and relevant that we do not see any discussion of the sort of tracking that is associated with socioeconomic class and degree of parental education in communities that are totally Caucasian. Why not? Should there be? Should we have court cases on this? Do we have court cases with regard to the tracking that clearly occurs in admission to our colleges? Why is it that many who are deeply concerned when the selection is *associated* with color of skin are not the least bit bothered when selection of similar differential intensity occurs inside the Caucasian population?

Jensen reviews what he regards as landmark court cases in Chapter 2. We suggest that the excerpts of judicial remarks do not lend confidence to judicial processes. We see the idea of innate ability and note that the word "innate" is used by all sorts of writers in various areas. We may surely regard some attributes of humans, such as color of eye or skin, as innate. But it is unreasonable, given what is known, to talk about innate mental or learning abilities, because we are then involved in the gene-environment question. We shall comment on this later. In the meantime, we suggest that the phrase "innate learning ability" be barred from the professional literature. We know that we can predict, not too badly, 12th-grade performance on a standardized achievement test from 6th grade performance. Does this make the 6th-grade performance innate? That we can predict suggests the possibility that some concatenation of circumstances led to the configuration of 6th-grade performance and that this leads with intervening environment to the configuration of 12th-grade performance. This is a *mere* statistical regularity of the past which tells us nothing about causality and which may or may not hold up in the future.

Discrimination. Jensen's Chapter 3 is entitled "The Drive for Equality." The term "discrimination" has come to have a pejorative meaning that is related to sex and skin color. Jensen says, and we agree, that discrimination is neither good nor bad. We need, for example, to classify humans on blood type - we discriminate individuals by blood type, and, of course, by tests. Obviously, with increasing specialization of human activities, and entirely appropriate demands by society for professional competence, we must have testing procedures to evaluate competence. And then because our schools and training institutes have finite and limited capacity, we have to develop procedures to form judgments about which individuals have successfully met the challenges of the specialized training. This is, of course, a matter of prediction. The area of specialization may be medicine, law, plumbing, landscape architecture, farming, cooking, and so on. Society must develop means of predicting degrees of success in *any* such direction.

What is the process of prediction? What is the logical structure of prediction? It is essential to consider this, particularly in view of the various writings on testing. The nature of prediction is nothing but estimating an unknown from knowns, the unknown being usually unknown because it is in the future. It could be in the past, however, as, for example, in forming a judgment about whether adult X had measles in childhood. To use a little mathematical symbolism: Suppose we want to predict a variable y for an individual for whom we know variables a, b, c. Then we have to develop a formula: $y = f(a, b, c)$ into which we can substitute known values for the individual for variables a, b, c and then compute f to get y. How do we get our predicting function? It is rather obvious even to the proverbial intelligent man in the street that we can do this only by looking at a segment of history in which we have individuals with known values of y as well as of a, b, c and then developing from this history, by data analysis and by statistical processes (which are actually very deep), a prediction formula.

Suppose we have developed a prediction formula. Can we say that it is the best possible prediction formula? We cannot. We can only say that this prediction formula is the best, on a stated objective basis, that we were able to find. We are led to wonder whether many critics of testing in general and of mental testing in particular simply fail to understand the process, highly empirical though it may be. In various applied areas, of course, the degree of validation of the prediction process is enormous - as with electricity, say, or with sending a space ship to the moon; but in other cases it may be rather doubtful, as with a nuclear facility or the prediction of the overall college GPA of a high school student. Considering prediction in this narrow (but correct) frame, we have to ask whether the prediction is valid for identifiable subgroups of the situations for which the predictions are being used. So, for instance, does the predictive equation derived from a sample of New Yorkers give valid results for Iowa farm boys? Or, of course, to turn to the nagging social question that permeates Jensen's book: Is a prediction equation developed for whites valid for blacks? It is utterly useless for writers to put forward the view that testing is unfair and inhumane and should be barred from societal processes (as some appear to do). The question of what we mean by "valid" will be discussed in detail later, but we can give a simple neutral example. We can ask whether an equation predicting heart condition at age 40, based on data at age 20 that has been developed from a body of history of whites, can be used on blacks, Chicanos, and so on. Obviously, such a question cannot be answered by pure thought. It can be answered only by statistical analysis of past data.

This view brings to the fore the point that a prediction equation is based on past history. We can easily imagine scientific or technological or social changes that would make prediction equations derived from the past not valid for the future. So, for instance, prediction of future performance of, say, black children may be quite erroneous if the circumstances of the predictions are strongly different from those of the past. We see essentially no recognition of this unassailable fact in the writings of Jensen and of the "hereditarians" in general.

Another totally critical aspect of prediction processes is that one simply *cannot* infer any sort of causality from the prediction equation we have obtained. To discuss this, we have to enter into very deep philosophical waters. If we were to regard causality as being merely high regularity of association of types of event, such as the association of water level on the Mississippi with the month of the year, we could regard the water level as being caused by the month, or even, to display the utter stupidity of the view, the month of the year as being caused by the level of water in the Mississippi. Why not, indeed? We take the view that it is epistemological nonsense to talk about one variable, such as heredity, causing another variable. IQ, unless we have done a massive experimental, *not observational*, study, in which we have systematically varied heredity.

Given data of individuals in terms of race and socioeconomic status (SES), family within race and SES, and individual within family, we can go through an arithmetical process called analysis of variance and then determine (to some extent, but usually with very binding limitations in observational studies) how much of the variability is *associated* with the *potential* explanatory factors race, family, individual and SES. Jensen (1980a, p. 43) says "race and SES contribute only 22 percent of the ... variance," using a common mode of expression (which we ourselves have used, erroneously). It is quite common, again, to say 22% of the variance is "due to" race and SES. Then it is common to translate "due to,' into "caused by," and then the fat is on the fire, as it should be, because *we cannot establish causation at all by the sort of study under discussion*. Jensen gives us (1980a, p. 44) a graph of IQ against SES for whites and blacks. Our reaction is simple! So what! The graph tells us what we observed. It raises interesting questions, perhaps. Should we ask: Does SES *cause* IQ? Does race *cause* a different association of IQ and SES? But these are, we insist, *stupid* questions. The variables IQ and SES are mere observational variables. If we could find a variable, that when altered, altered SES, we could say that that variable is one of the causes of SES, and similarly for IQ. If we want to see whether SES causes IQ, the *only* thing we can do is to find a controllable variable that changes SES and see whether it causes IQ. We are writing all this at a little length because the matter we discuss lies at the root of all the nature-nurture-IQ controversy. We find an absence of this type of thinking in Jensen's book. We believe that we are not misinterpreting Jensen to say that he exhibits little perception of the nature of causation and of the intrinsic difference between correlation (or *regressions*) and causation (in spite of brief caveats he gives).

Jensen does consider the question of prediction. There is (1980a, p. 47): "the technical problem of establishing that the predictor is in fact substantially predictive." This, obviously, pervades the whole business. In Jensen's context, and naturally so, prediction must relate to performance. On the question of bias in mental testing, Jensen says (1980a, p. 48) "bias exists when the method of selection discriminates individuals differently than does the criterion measure of performance." We cannot fault this as a general statement on one critical aspect of bias. But we have to analyze it. Obviously, a very basic question concerns the validity and appropriateness of "the criterion measure of performance." In connection with the use of IQ testing, we have to ask what are appropriate criterion measures of performance. We found Jensen's writings weak on this, in that the real difficulty of finding appropriate criterion measures is not discussed.

Mental ability. Obviously, in the context of Jensen's volume, we have to ask what mental ability is and whether we can measure it. What has happened in the mental testing area, we suggest, is that the development has been entirely empirical with little discussion of the basic question of what mental ability is. The question is obviously very difficult. Anyone who has taught realizes the existence of individuals who are "bright" and individuals who are not. One can see the phenomenon in three-year old children. The processes followed by mental test constructors are described extensively by Jensen in his Chapter 4, "The Distribution of Mental Ability" and Chapter 5, "Varieties of Mental Test Items." These two chapters exhibit, we believe, the very weak status of an (admittedly) incredibly difficult area. What has happened, it appears from Jensen's description, is that the whole process consists of making up test items, then testing these for some purely statistical coherence, and then examining the total score that a battery of test items gives. The basis of this process is the *assumption* that there is such a thing as "general" (our word) mental ability and that the measure of this should be distributed according to the Gaussian (normal) distribution.

First, Jensen talks about the distribution. Why have testers "settled on" the normal distribution? Jensen's answer (1980a, p. 72) is:

> The simple fact is that a test unavoidably yields a near normal distribution when it is made up of (1) a large number of items, (2) a wide range of item difficulties, (3) no marked gaps in item difficulties, (4) a variety of content or forms, and (5) items that have a significant correlation with the sum of all other item scores.

Significantly, Jensen says, "items that are uncorrelated or negatively correlated with the total score can only add error to the total scores" (p. 73). This bland statement gives us great pause. Its obvious implication with regard to the process of making up a battery of test items is not discussed. For instance, if we have 20 items that anyone would regard as indicating some sort of mental ability, then if another proposed item does not correlate significantly or even correlates negatively with the total score for those 20 items, should it be thrown out? What does this imply about the whole of the achieved battery of test items? Suppose the 20 items reflect, say, arithmetic ability. Then arithmetic ability becomes "king," and other types of ability will not be allowed into the battery. There are, we believe, deep philosophical issues involved here, but we do not find them discussed. To discuss them is very difficult to be sure, but in our opinion Jensen opens the issues and goes nowhere.

Actually, however, we believe the constructors of tests are indeed aware of the difficulties. Have they resolved them properly? That, we judge, is a good question with respect to the IQ test. In the case of the SAT (scholastic aptitude test), we do have verbal and mathematical portions. But in a single IQ test, some judgment must be made of the frequency of types of question in the whole battery. How is this to be decided?

Then, of course, we know that the general tests do not examine a variety of mental activities, like musical, artistic, and cooking ones, at all. They are not designed to do this. But we may well call into question the exclusion of a very significant portion of the mental abilities that contribute to the "good society." The summit of this line of thought is that we must reject the idea that we can map the whole range of human mental abilities onto a single number line. This has been a standard criticism of the IQ test, and we judge it to be entirely valid.

But are we to go to the other extreme and say that there is an infinity of distinct mental abilities? To do so would be useless from the viewpoint of the individual or of society. The test constructors have tried to construct measures of the mental abilities that are common to a very wide variety of vocations and have been concerned with criterion validation within vocations. They are not to be classified as testing gnomes who are foisting their own prejudices on the outside world.

Then we get into one of the very basic difficulties: "the scale problem" (1980a, pp. 74-75). Jensen wants an interval scale for various reasons, but then says, "unfortunately we have no direct way of knowing whether the scores on most mental tests constitute an interval scale" (p. 74).

On pages 71-95 are some weak arguments for interval measurement of mental abilities. We offer as examples the following statements: "Scientists have never argued about the distribution of height, or brain weight, or life span, or pulse rate, or the air capacity of the lungs. So why should there be any dispute about the distribution of mental measurements?" (p. 74).

> We simply *assume* what the distribution of scores should look like if we had an ideal test that measured the trait or ability in question on a perfect interval scale. Then, 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 - provided, of course, that we are correct in our initial assumption about the true shape of the distribution. (1980a, p. 75).

The answer to the quoted question is that there is only one scale for measuring each of such variables as height, brain weight, and life span. These scales are dictated by the laws of physics, where physical laws are based on well-supported theory. No such theory exists in psychology and, according to Jensen, "it is claimed that the psychometrist can make up a test that will yield any kind of score distribution he pleases. This is roughly true" (p. 71)! We offer other Jensen quotes (pp. 95-96): "There is good reason to believe that achievement, in contrast to more elemental traits and abilities, is not normally distributed in the general population but that it has a markedly skewed distribution," and "Most scholastic achievement tests, however, are constructed in such a way as not to reveal the skewed distribution of achievement." We assert that the observed distribution of test scores cannot reveal the underlying distribution of the theoretical construct for abilities or achievement. Jensen cannot have it both ways. He cannot say that the empirical distributions of ability test scores do reveal the underlying distribution but, because of the way in which achievement tests are constructed, they do not reveal the underlying distribution. Further, it would not be difficult to construct two variables that were moderately related, with each variable normally distributed but with the relationship between the two variables not linear, indicating that the interval sizes were not the same for the two variables.

The idea, the *hypothesis*, that there really is a general measure of mental ability pervades Jensen's exposition of test construction. The items in a battery *must* have positive interitem correlations. Jensen says (1980a, p. 69): "Without positive inter-item correlations, test scores would represent only error variance." This, we have to say, does not make sense. It is *petitio principii*. What indeed is "error" in Jensen's statement? There are special circumstances under which we can formulate a reasonable concept of error. For example, consider a muscle strength measurement, such that we can measure it for an individual many times. Then we can consider the variability between the measurements as "error." But in the case of a collection of test items there are problems. We can, of course, imagine giving a large battery several times to an individual with the *hope* that there are no memory effects, and we would then get something that we could reasonably call the "error" of an item.

We see the kind of error (made initially by Galton, one of the heroes) that involves attaching almost divine significance to the normal law: "Because errors of measurement are by definition random and independent... they are distributed according to the "laws of chance," which means that they are distributed 'normally,' that is, according tó the normal curve." (Jensen 1980a, p. 70). This is plain rubbish! It is true that if we have many measurements, each independently subject to error (with finite variance), then the mean will be distributed somewhat like the normal distribution. The absurdity of calling on this mathematical theorem is exemplified by considering the average of a set consisting of 20 measures of physical ability, 20 measures of mental ability (e.g., by arithmetic tasks), 20 measures of hearing ability, and 20 measures of personality outwardness, each on a scale of 1 to 10. One will find that the overall average is distributed rather like a normal distribution. One should merely react, so what! Then again, Jensen calls on the empirical fact that physical measurements, such as those of height, show "approximately normal distributions." This occurs if we choose as our measuring stick the usual one, and is a very interesting and curious experiential fact. But what allows us to suppose that this lends force (not mere suggestibility) to normality of distribution of mental ability, whatever that is?

The question of whether IQ test scores give an interval scale bothered Burt (a rather profound thinker, in spite of some very odd behavior). We see (Jensen 1980a, p. 75): "*Ipso facto,* any test of intelligence that yields a normal distribution of scores must be an interval scale" We regard this as rubbish. Burt as well as Jensen (1980a, p. 79) justifies this on the basis of the polygenic theory. Jensen says (p. 80): "The polygenic theory of individual variation in mental ability leads us to expect a more or less normal distribution of ability in the population." This theory is nowhere near as powerful as Jensen thinks. This is not the place to go into the theory of quantitative genetics. However, it is worth stating that the critical basic assumption is that the attribute considered is made up *additively* of many small random components that are independently distributed and associated with the genes with an additive random environmental contribution. Then, under additional assumptions on mating and Mendelism, one can obtain theoretical expectations of correlations between relatives of various degrees. It turns out that the observed correlations correspond moderately closely to theoretical expectations, which have, of course, parameters to be determined by the data available. Can this outcome be regarded as validation of the associated theory? Jensen and other hereditarians clearly hold this opinion, and strongly so. This opinion, with associated so-called heritabilities of 80% (or even of 60%), must be faulted on several grounds: (a) the theory includes assumptions that are *patently* false; (b) the process that is followed is solely that of observational science, that is, observing correlations in an existent population, but then interpreting the theory as a causal theory - as though in some future scientific world, we shall be able to quantify and experimentally control the genetic variables and the environmental variables, and when we do so we shall find the *interventional* causal results that the observational processes suggest.

On the all-pervasive use of normality Jensen says (p. 87): "Finally, psychologists accept the idea that intelligence is normally distributed because no compelling alternative theory or evidence for any other kind of distribution has ever been proposed." So we have an almost mystical belief. And to increase our discomfort with this appallingly weak defense, we note that the test constructors (whom we do not hereby castigate) have constructed their tests *so that they do give at the end distributions that are approximately normal.* The circularity of the process is surely entirely obvious. We do not wish to seem to be expressing the view that test constructors are stupid. But we insist on our view that the whole business of quantifying intelligence is a bootstrap operation (as is all science, incidentally), and one cannot use *petitio principii.*

The generally accepted theory of mental testing is permeated with the basic mathematics of normal, univariate and multivariate, distributions. Naturally, then, Jensen gives a (largely verbal) account of this. It is sufficiently abstruse to give mathematically untrained individuals difficulty and to give such individuals an impression of "high" science. But to a trained statistician, it is "small potatoes" and does not possess any revelations that must be taken to be compelling.

Jensen tells us about the well-known IQ distribution in black and white populations in the United States. The test is normed on whites with a mean of 100 and a standard deviation of about 15. One finds that the mean black score is, say, about 85, with a smaller standard deviation. We see again and again criticisms to the effect that this is a product of some bias in the test procedure and scoring. Our own view is that one can be highly critical of the IQ test as a measure of intelligence, but there can be no question that the test is objective, and the results, unpleasing as they are, and presenting serious social questions as they do, cannot be regarded as racially motivated measurement. And as Jensen describes, and we discuss later, the IQ test scores have some degree of predictive ability for attributes that are socially useful. IQ test scores have been used to classify children with respect to retardation. This gets into very deep waters. One can readily understand how parents are considerably upset by the discovery of retardation; this happens with highly educated and able parents, but more often with less educated parents of the various possible colors. Is this classification to be rejected as an invidious labeling by evil testers? Obviously not, we assert. The test score is an experiential fact that is clearly individually and socially useful in forming assessments of children. The IQ test, however weak its scientific basis, was developed by Binet for such purposes, and some analogue of it will be needed for the indefinite future. It is curious to see writings by academics who castigate the whole process but who teach bodies of material and then test their students for their ability to understand and to use them - surely a testing of a component of intelligence, whatever that is. Jensen gives us his picture of how IQ is *associated* with achievement, and we must, we believe, accept the general picture he gives.

However, when we turn to Jensen's exposition we are less comfortable. On standardization of testing, we find (1980a, p. 126):

> An essential part of the meaning of "standardized" is that the stimulus or situation eliciting the behavior that is to be observed, rated, or graded should be relatively unambiguous and objective, in the sense that it is perceived consistently as the same task by all persons and by the same person at different times. It should present no choice and no difficulty in terms of the subject's knowing what he or she is *supposed* to do.

This is a very fine platitudinous statement, and we suggest that the reader, whom we imagine to be college educated, run through the test items that Jensen gives in Chapter 5 and judge whether the requirements are met. Then we see on page 127: "in an intelligence test the specific content of the items is unessential, so long as it is apprehended or perceived in the same way by all persons taking the test." It seems crystal clear that the test items on pages 148-150 involve two aspects, *actual knowledge of the language in the items and also 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 or low scholastic aptitude or whatever? It is surely justifiable to use such items to develop a test of language comprehension. But Jensen does not, it seems, discuss this. These items would lead to the "atrocities" of the early part of this century in which individuals of, say, Russian origin were labeled as morons. If a test battery of such items is used to tell parents that their children are so poorly familiar with the language use of ordinary schools that they need to be placed in special classes or schools, then it is hard to see how a rational criticism can be mounted. Does Jensen appreciate this? We surmise that he does, but then we ask, Why is there no discussion?

Analogy test items are used. It seems obvious that the making of analogies is partially a trained ability. Obviously, given knowledge of the language and wide exposure to the making of analogies, the making of correct analogies is a mental ability. It is interesting to look at science, say physics, and see the tremendous role of analogy and the way brilliant people have made useful analogies that subsequently proved to be utterly fallacious. The making of analogies is a deep process, and if youngsters have not been widely exposed to doing so, they will be unable to do it with test items. Such individuals may justifiably be labeled as *environmentally handicapped,* but to assert that they have *inherently* low mental abilities is rather shocking.

Next we have pictorial tests. Given a group of children with the same environmental background, we are sure to find differences that correlate with differences on other test items. But the child who has not been exposed to such problems from an early age will perform poorly relative to one who has been. Can we use this poor performance to infer low "innate" ability (whatever that is!)? Of course not!

It is easy to respond to the points we are raising by saying that it is a favorite sport of antitesters to pick out single items and criticize them. And, of course, *and unfortunately*, it is all too easy to ridicule testing in this way. It is desirable, then, to make some remarks about "strange" or "weird" items in test batteries. This is a common route of testing detractors. A statement by Jensen that we wish to endorse is (1980a, p. 128):

> One must realize that no single test item is a very good measure of intelligence... This fact is in large part the basis for the plausibility of those criticisms of IQ tests that consist of singling out specific items as examples of the supposed triviality of what is measured by the test.

This point is important because it illustrates that lay persons cannot sit back in their armchairs and level cogent criticisms at testing or any other highly complex field. Even though experts are fallible, we must depend upon them to a large extent. It is unreasonable to assume that the average expert is biased and will use data to support these biased views. Individual experts, as well as lay persons, find certain items silly, but neither the experts nor the lay people will agree on which items appear silly. Experts, for the most part, will revise an opinion about an item they regard as silly on the basis of a large amount of data indicating that the item is functioning well. The lay person does not have access to such data and would not have the expertise to evaluate it. We do not wish to imply that experts should ignore the criticisms of the public, but we find it unreasonable that certain members of the public persevere in such criticisms despite the fact that these criticisms have been attended to by experts and appropriately answered by them.

Do IQ tests really measure intelligence? At one time, Jensen said, we believe, "intelligence is what intelligence tests measure." Jensen (1980a, p. 171) quotes Wechsler:

> "What we measure with [intelligence] tests is not what tests measure.... These are only a means to an end. What intelligence tests measure, what we hope they measure, is something much more important: the capacity of an individual to understand the world about him and his resourcefulness to cope with its challenges."

This is surely a reasonable homily. But it leaves us with much uncertainty. And, also, we call to mind individuals who undoubtedly scored very highly on IQ tests, yet are rather hopelessly unable to cope with the challenges of the world. The statement of Wechsler does not tell us what (intelligence) tests measure; it tells us what he, for instance, hopes the tests measure, but it also opens up the question of what challenges the subject is to cope with. This, of course, brings in the question of predictive validity which Jensen takes up later. We see, rather frequently, the question that can be written as, Is intelligence a "thing"? (cf. Gould 1980). Intelligence is a construct. It is not a "thing," and no one, not even Jensen, believes it is. In the same way, temperature is a construct that occurs in various forms in our ideation. In mathematical physics it is a mathematical entity that enters into various mathematical models. In the real world, it is what a measuring process gives, and we have conventions for standardizing the measurement gadget and process.

It is all too easy, however, to interpret some writings in the field as asserting that intelligence is a "thing." This leads some to say, then, that IQ becomes reified. The appropriate interpretation in the whole area is, we suggest, that while it is hard not to construe some writers as reifying IQ, the reification is not accepted by the great bulk of professional testers. There is, however, the experiential fact that IQ and scholastic achievement are highly correlated; and it is not, surely, a defect of a society to attach high value to high scholastic achievement. Also, because IQ is predictive of such achievement, it is easy to make the silly logical mistake that IQ *causes* degree of achievement, and then reification becomes correspondingly easier. Our opinion is that Jensen comes very close to reifying the general factor.

Jensen takes us to a section entitled "Armchair Analysis versus Empirical Investigation," and gives us Intelligences A and B and C. He questions the appropriateness of our criticism in our first-round review (Kempthorne & Wolins 1980). Given a test that is taken to measure intelligence, the actual score that an individual gets on a certain day in the particular testing environment is the phenotypic intelligence, Intelligence C - at that time and in those circumstances. Then Jensen tells us (p. 184) that Intelligence B is the individual's general intelligence which "cannot be properly understood without some basic conception of factor analysis, from which the notion of general intelligence gains its scientific meaning." (So then we are faced with the necessity of validating the ideas of factor analysis, to which we shall turn later: We shall see great difficulties.) Intelligence C is then a score on a particular intelligence test. We certainly cannot argue about this. We have a test, with a scoring procedure, and the application of this to an individual gives a score. There is, then, no such single thing as Intelligence C. If we have m "intelligence" tests $T_1, T_2, \ldots, T_m$, we have Intelligence Cs, say $IC_1, IC_2, \ldots, IC_m$. Then, we have to follow the ideas of factor analysis and extract Intelligence B. Even if we accept this process, how are we to get Intelligence A? In more detail, suppose we have accumulated, by hook or by crook, our m tests, and have $IC_1, IC_2, \ldots, IC_m$ for a large number of people; suppose, then, that we have an arithemtic procedure to obtain a formula.

$$IB = f(IC_1, IC_2, \ldots, IC_m)$$

Then we can for a "new" individual obtain $IC_1, IC_2, \ldots, IC_m$ and apply our equation to obtain IB for that individual. We can accept this as an objective process to obtain IB, though whether this should be given the name Intelligence B is not clear except by fiat. Now we have to construct Intelligence A (genotypic). To get this we have to use ideas of quantitative genetics (which we mentioned a little earlier). We can get a grip on this with species that we can manipulate genetically and environmentally, but in the human intelligence area intelligence A is an entirely unrealizable construct. We can, to be sure, use a statistical process with more or less genetic modeling to make an "estimate" of it. But at this point the status of intelligence A is so murky as to be, in our opinion, useless with regard to the science of the mind.

Finally, we noted the definition of Humphreys (1971); (Jensen 1980a, p. 170): "'Intelligence is defined as the entire repertoire of acquired skills, knowledge, learning sets, and generalization tendencies considered intellectual in nature that are available at any one period of time.'" If we accept this as an *informal* definition - and we think that we should - it tells us without a shadow of doubt that intelligence, and hence the "g" we locate, are not something innate, and determined largely by genes. Passing from this informal definition to the belief of some, including Jensen, that intelligence tests capture an intrinsic unvarying (with age and environment) attribute is *obviously* not warranted.

Factor analysis. Jensen calls strongly on factor analysis for his definition of intelligence. The books on factor analysis that we are familiar with (Harman 1967; Lawley & Maxwell 1963; Mulaik 1972) are confined to the mathematics, numerics, and probability aspects. Jensen confines himself to the principal component analysis of the reduced correlation matrix (the diagonal unities being reduced to communalities) and then focuses his discussion almost entirely on the first principal component that is obtained. This is an unreasonable choice, given Jensen's overall aims, because the loadings on the obtained general factor reflect in part special abilities in addition to the target construct, g. There are great difficulties that are not discussed.

To explain the difficulties is not easy, and understanding can be achieved only by understanding the mathematics and numerics of what is done. Suppose we have a large battery of test items. Do these items reflect "intelligence"? How are we to decide? We shall follow, let us say, Jensen's procedure and obtain the principal components, say, C_1, C_2, . . . , C_k of the reduced correlation matrix. How are we to pick out items that reflect intelligence? In general, we shall have to do what is called a rotation, and then we shall designate one of the rotated factors as the intelligence factor. The whole process is profoundly obscure. That this is so is not surprising. Fine minds, Spearman, Burt, Thurstone, and Guilford, to mention a few, have contributed to ideation on the problem. There are considerable difficulties with respect to distinction between the covariant and specific parts of variables, as we see from the following quotations from Mulaik (1972): "But, as we shall see, not all factor analysts agree on the formulation of this distinction, some claiming that the model is vague and indeterminate" (p. 133); "Thus the model of common-factor analysis is indeterminate" (p. 135); "The basic problem of common-factor analysis is the determination of the unique variance of a variable" (p. 135); "he [Thurstone] may have overlooked some difficulties for the generalization of his results" (p. 135); "In the preceding chapter we discussed the model of common-factor analysis, pointing out that its chief defect lies in the indeterminacy of the common and unique portions of variables" (p. 173); "Hence there is an indeterminacy in determining the common and unique factors from within the common-factor model" (p. 327); we suggest careful reading of the whole of Section 133; "But indeterminacy is not confined to just the $(r + n)$-dimensional space encompassed by the model. Implicit . . . is the existence of an even larger space of variables and 'factors' in which the observed variables are embedded. This is the space of all possible variables which may be defined logically on a population of interest with respect to a domain of attributes" (p. 327).

In providing these quotations, the sense of which we agree with, we are not being derogatory of psychologists or of factor analysis. The whole effort is a fine creation of human ideation. But our quotations, if we accept their approximate truth, tell us that anyone who bases his theory on the general factor is on very uncertain ground. Jensen calls on g, the so-called general or common factor heavily. He gives us a section "The Nature of g." He discusses "the problem of domain and the uniqueness of g." Some of his discussion is enlightening. We noticed, however: "It seems a safe generalization that the g of a *large* and *diverse* set of mental tests is the same as the g of a different large and diverse set of mental tests" (Jensen 1980a, p. 233). A natural reaction is obvious. From Mulaik's remarks quoted above, for example, it does not seem a "safe generalization." Indeed, it seems, to us, an *unsafe* generalization.

We shall not discuss the remainder of Jensen's chapter, which is of some interest. But we must quote page 249: "A working definition of intelligence, then, is that it is the g factor of an indefinitely large and varied battery of tests." We suggest that this lacks precision to an unacceptable degree.

Jensen is obviously in a very difficult area, and we sympathize with him. Most of the books on factor analysis do *not*, in our opinion, address the substantive issues significantly. Jensen *does* attempt to do so, but we would have liked more exposition and discussion. The "logic" appears to be as follows: One looks at the principal components of the reduced correlation matrix. One then looks at the loadings of items on these components. One makes a *judgment* (which is certainly not unfounded) about which items reflect intelligence, and then one searches for a rotation that loads "strongly" on those items. One then calls this factor achieved by rotation, *the* general mental factor. Next, in the search for other items that are to reflect "the general intelligence factor," one selects those items that correlated strongly with those items that have large loadings on "the general intelligence factor." This, it seems, is what is done in constructing a test battery according to the ideas of factor analysis. The procedure surely seems complex and is *not* understandable unless one comprehends, at a fairly deep level, the mathematics of factor analysis. Should we accept the whole of the recipe? Jensen tells us (p. 215) of Thurstone's seven primary mental abilities and the fact that each of Thurstone's subtests "measures g as much as, or even more than, it measures the particular primary ability." He describes Eysenck's factor analysis of 60 of Thurstone's cognitive tests, which exhibits a g factor and other analyses that do likewise.

A deficiency we noted in Jensen's presentation (which is also, we judge, a deficiency of many texts on factor analysis) is in the discussion of estimating the general factor levels of the individuals in a correlational study. We are told that the g factors of different test batteries are strongly correlated, but the technique of addressing this facet, and then the results of using this information to examine whether the Stanford-Binet and the WAIS, for instance, do give g's that are very highly correlated, are not presented. This is surely necessary if we are to accept the proposition that different batteries do in fact give nearly the same g. Jensen tells us (p. 223) that "total scores on a test of many g-loaded items will order individuals in about the same way as the individuals would be ordered in terms of their g factor scores." We would like to see "large" evidence on this.

Although it is obvious that we do not find the evidence Jensen gives for the ubiquitousness of g complete and compelling, we have to state our opinion that the evidence is strong. The existence of a general factor is not an artifact produced by the testers with the aid of arbitrary numerical processes. Opponents of the idea must produce objective, not merely emotional, objections and must produce an alternative explanation of why objective tests and objective data-analysis procedures produce the regularities that Jensen describes. Then, in addition, we must say that even though the construct validity of a general factor has not been demonstrated, the whole of the ideation has produced, as we see later, predictive devices of great utility. The IQ test is integrated with g-factor thinking and is predictive. The nature of the mechanisms that produce the outcome, genetic and environmental, is quite unknown. All in all, the results suggest to us that the concept of g needs further clarification. It is a promising construct which may enhance our understanding of human mental differences. We need to develop tests with greater diversity of content, as Jensen implies, and to use these to evaluate research directed at understanding the environmental and genetic mechanisms that produce the observed variation.

Our own opinion is that the evidence against the null hypothesis of there being no general intelligence factor is very strong. However, the characterization and quantification of such a factor are extremely tenuous. It is surely the case that Jensen's procedures overestimate the variability associated with such a general factor.

We suggest that to regard Jensen's writings to this point as justification for what goes on in the intelligence testing area is a major mistake.

Reliability and stability of mental measurements. Here we meet the individual's obtained test score X, the individual's hypothetical true score T, and the individual's obtained score on some criterion C. There can be, it seems, no question on the first and last, but the middle construct occasions real difficulties. Reliability, relevance, and validity present serious problems, of which Jensen is obviously aware. On reliability, the most primitive and acceptable idea is assessment by the "split-half" technique, but, *of course*, this is only a means of estimating the reliability of a score from the whole battery and not a means of estimating the hypothesized and very vague totality of test items that are used to quantify intelligence or mental ability. The person who is not familiar with the whole folklore of test construction by professional test constructors will be skeptical about the concept and operational quantification of reliability. However, our judgment is that in spite of the problems of definition and measurement of reliability, the techniques used in the field are very reasonable in the light of the vast difficulties of quantification of mental activities. We must accept Jensen's tables of reliabilities (pp. 271-73) as having considerable force as *experiential* facts.

But we surely must recognize that Jensen is in deep trouble by the time he gets to Table 7.6 (p. 279). Here we see the intercorrelations of Stanford-Binet IQs at various ages. We see that the correlation of IQ at age 3, say, with IQ at 15 is 0.43. Jensen gives the first principal component (PCI) in this table and is highly satisfied that this comes out to be in the range .71 to .93. Then he says (p. 278): "This general factor accounts for 77 percent of the total variance in IQ between the ages 2 ½ and 17 years." Also (p. 279): "The general factor in Table 7.7 accounts for 86 percent of the total variance in all test scores between Grade 1 and Grade 6." Is Jensen entitled to take comfort from these statistics? They are like what we observe with height and weight (Table 7.8, p. 280). We react: So what! We do not regard this evidence as forcing in the way Jensen does. Instead, we look at the quality of prediction of IQ at age 17 from IQ at age 3, and we do not like what we see (in relation to classifying children, e.g., as educationally mentally retarded). Can we regard IQ at the various ages as we regard scores on different tests at the same age - to which we may, with some hesitation but with some justification, apply factor analysis? One can *always* apply a particular numerical technique to a data set of particular form. This is, however, mere data analysis, which may be suggestive, but has no higher status. One should, however, look at this table from other viewpoints, such as a Markov process one. The factor analysis model is nonsensical in this context.

Validity and correlates of mental tests. Validity is, of course, an essential area. Testers have brought in four ideas concerning validity: content validity, criterion validity, concurrent validity, and construct validity. In the general area of performance tests, there is obvious need for content validity; tests must measure "some clearly defined universe of knowledge." The underlying problem, not addressed by Jensen, is the question of content validity of IQ tests. We saw very little on this. Criterion validity refers to the ability of the test to *predict* performance external to the test. The only problem is to quantify the outside tasks reasonably. We then become involved, with IQ testing, in what these are, and this is very *difficult in respect to general education. This is the real* difficulty. Finally, construct validity is "more difficult to explain" (Jensen 1980a, p. 303). Indeed it is. What "really underlies the test"? Our impression is that the construct validity of IQ tests is of dubious extent - for the simple reason that we have made little progress in determining what intelligence is or should be defined to be. The whole area of the validity of mental tests is still, we suggest, very obscure.

When we turn to correlates, the story is just a matter of statistical calculations. Jensen says (p. 313): "IQ has more behavioral correlates than any other psychological measurement." We accept this as an experiential fact. Undoubtedly, there is a bias in research in that the IQ test has been examined in this respect much more than other mental tests. The "bottom line" is, at least in connection with education, the correlation of IQ with scholastic achievement. There are, of course, really serious problems here. How are we to measure scholastic achievement? Can we take overall GPA? Of course not. We guess that this is relatively easy in primary grades, with nearly complete emphasis on reading, writing, arithmetic, and memory and the understanding of factual information. Beyond this point, there is an obvious tracking of youngsters into streams of varying mental depths. The overall GPA by the end of high school can be an absurd measure of scholastic achievement. This aspect is further magnified in college, where one finds highly graded students in some subjects who cannot write a reasonable expository short essay and cannot add and multiply fractions. However, the experiental fact is that GPA does correlate with IQ very appreciably, even though, to quote Jensen (p. 331), "GPA per se is a poor index of actual achievement."

A relatively recent incursion into educational and other social studies has been the use of path analysis [developed by S. Wright (1921) and *proper* to some *theoretical* genetic situations]. Jensen says (1980a, p. 336): "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." Dorfman (1980) questioned this in his review, and we also fault it, on at least two grounds. In the first place, it is, we believe, *scientifically* quite unsound even to talk about one observational variable causing another in the absence of experimentation (as we discuss above). In the second place, path analysis is a method of quantifying the "size" of paths in a *given* path diagram, with an assumed linear structure of relationships among observed variables. One *cannot* infer causal relationships in the useful sense of determining the effect of experimentally induced changes in supposed "causal variables." One gets out of path analysis, at best, just a little more than one puts in the diagram. Throughout Jensen's discussion runs his attempt to discuss questions such as IQ causing occupational status or income level. We reject this mode of expression completely. Incidentally, the quotations on path analysis given by Jensen (1980b) in his reply do not conflict with what we say, though they do not spell out what really is being achieved.

Test bias. By now, we get to the main topic of Jensen's book. We are told (1980a, p. 367) that Binet "fully recognized that language, cultural background, and a common background of experience were necessary *vehicles* for the measurement of intelligence." Apparently Jensen agrees, and we are then forced to accept the view that our various national minority groups do have "language, cultural background and a common background of experience," as do our rural and urban, male and female sexes, our children from poor white families, and so on. That this is so patently false is, perhaps, the main reason why so many writers reject the claims of the hereditarians and the proponents of the IQ test as revealing the "real" extent of intelligence in individuals.

Jensen discusses cultural bias in test items. His own exposition tells us that the test items reek with cultural bias - *children will not* "know and be familiar with the subject matter or specific processes required by the test item." Jensen, we judge, passes over all this, and turns to his definition (p. 375): "In psychometrics, bias refers to systematic errors in the *predictive validity* or the *construct validity* of test scores . . . that are associated with the individual's group membership." Does the exposition meet this definition "head on"? With respect to predictive validity we are given (p. 381): "*A test* [with perfect reliability] *is a biased predictor if there is a statistically significant difference between the major and minor groups in the slopes* b_{YX}, *or in the intercepts* k, *or in the standard error of estimates* $SE_{\hat{Y}}$ of the regression lines of the two groups." (Y is the criterion score and X is the test score.) Jensen discusses the underlying statistical ideas, which are not at all trivial, rather well, we think. Can we apply this definition? Does Jensen apply it? We have to say emphatically, Yes! We see a satisfactory (at the chosen level) exposition of the theory of the situation, with a very extensive discussion of various aspects and of the literature. Jensen claims (p. 515) that "in the vast majority of studies, the regressions of criterion performance on test scores do not differ for blacks and whites." We are strongly inclined to accept this conclusion. We see no bias in Jensen's presentation, and he offers, we think, an honest account of the situation.

The only lingering question in our minds is that of construct validity, which (p. 420) is "a complex, open-ended affair." We surmise that many readers detect a critical weakness in this respect, which influences strongly their overall reaction to the book. We are given (p. 421) Humphrey's statement with respect to blacks and whites: "'*While there are obvious environmental differences, these differences are not so profound as to require different psychological principles in the explanation of black and white behavior.*'" We balk at this, because he offers this as a statement and not as a judgment. Jensen calls on the similarity between races of the correlation of raw scores and age. Such a similarity is necessary, but not at all sufficient. He calls on kinship correlations, and the same view must apply. Then Jensen calls on groups-by-item interaction, and regards the (essential) absence of this as strong evidence for construct validity. This is indeed evidence that the test is measuring the same construct in the two groups. But if the construct that the test aims at is useful in whites and in our present society (as seems unquestionable), then the evidence indicates that the large societal problem we have is not an invention of the test constructors. Factor analysis of test items gives further reinforcement but gives no evidence on the reason for the white-black difference.

We judge that Jensen has provided a competent examination of empirical evidence on bias in predictive validity, 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 measuring body weight is not biased because it gives, say 90 pounds as the weight of a 21-year-old male 6-feet tall, when we know that the weight of such a person in normal circumstances would be about 175 pounds. The measurement process is telling us what we need to know, the weight of the man. It does not tell us, of course, the causation of this very low weight.

One lingering question that must give us some pause is the aptness and quality of the criterion performance measures. We are told (p. 472) that the published evidence with regard to IQ and scholastic achievement in elementary school is "surprisingly meager" and in high school (p. 474) is "surprisingly scant." The validity of SAT scores for predicting academic performance in college is discussed by Jensen. The outcome is that there are trivial differences in quality of prediction for whites and blacks. If we accept the criterion performance variables that were used (and this seems reasonable) we have to conclude that (p. 515) "most standard ability and aptitude tests in current use... are not biased for blacks or whites with respect to criterion validity."

Criteria of test bias: Empirical evidence. The general message of Jensen's chapter on empirical criteria, which we judge to be quite well-supported, is that identifiable portions within the test battery "hang together" within racial groups. That is, the correlations between tests are essentially the same *within* groups. The differences between racial classes in performance on items that might be judged to be culturally biased are of essentially the same magnitude as with items that seem to be less culturally biased. Also, perhaps curiously, white-black differences are generally slightly larger on nonverbal than verbal items, whereas one would expect the reverse if language ability reflected whatever differences in culture there are. But we must say that there are scaling problems that could modify this naive interpretation.

Jensen claims that factor analyses of test batteries in white and black samples show the same factorial structure. This is a necessary condition for validity of comparison of means, but, of course, not a sufficient one.

On page 587, in the chapter summary, Jensen says: "All the main findings of this examination of internal and construct validity criteria of culture bias either fail to support, or else diametrically contradict, the expectations that follow from the hypothesis that most current standard tests of mental ability are culturally biased for American-born blacks." Our judgment is that Jensen has in fact given an essentially correct report, reasonably and usefully addressing the basic issue.

External sources of bias, sex bias, culture-reduced tests and techniques. We judge that the material of Jensen's chapters on external bias, sex bias, and culture reduction is a very reasonable report of research and thinking. The upshot is that the first two factors do not lead to significant problems. Jensen also gives, we believe, a reasonable account of the third item. An interesting result is (p. 713) that "such culture-reduced tests do not show smaller mean differences between blacks and whites (in the United States) than do conventional culture-loaded IQ tests." This really does seem to be the case. The reader and interpreter of the ideas and issues here must appreciate properly the effect of differences in reliability without which an incorrect interpretation can easily be made.

Uses and abuses of tests. After his argument that the IQ test is a valid measure of a supposed g factor that really quantifies intelligence, and that it possesses the various types of validity, one would expect Jensen to take the position that IQ testing is useful. Instead, Jensen sees (p. 738) "no routine purpose... that cannot be better served by standardized achievement tests." We may wonder why there is all this writing about IQ tests. Why bother with them? We find it curious that Jensen seems to accept standardized achievement tests without significant discussion.

Jensen closes his work with various statements that surprise us: "Ability grouping at the elementary school is more a convenience for teachers than a benefit to the pupils" (p. 738). We ask: Is it reasonable to have in a 5th-grade class students who are at grade 10 and at grade 2 level in reading? After all the supposed scientific validation of testing, we read (p. 740): "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." But Jensen, in this book and in other writings, has, clearly, made very strong efforts to do so. In other writings, it was the genes that caused the differences; in this book it is the differences in the amount of the g-factor. On minimal competency testing (MCT), Jensen says (p. 724): "I cannot see MCT as in any way contributing to the solution of the problem." The problem here is the high failure rate (e.g., in Florida) with regard to arithmetic, reading, and writing. Our reaction is that no one has suggested that MCT contributes to the solution of the problem. It surely demonstrates the existence of a problem. Also, routine year-by-year use of MCT-type tests would exhibit the immanence of the problem of the high school diploma being a false certification of basic abilities.

An issue of fairness. Jensen (pp. 493-94) raises a large problem associated with ideas of predictive accuracy. Suppose we accept, say, the General Classification Test (GCT) score and final school grade, and the appreciable predictive value of the former for the latter. Then, accepting predictive accuracy as the critical aspect, Jensen tells us that we should use separate predictive equations for whites and blacks, in other words, include race as a moderator variable (p. 494). We have to take the view that this should not be done, on the basis of fairness. The group membership of an individual should not be used in predicting criterion (e.g., job) performance. We are quite unable to justify this except to give our opinion that selection of individuals must be based on individual merit and must not make use of group membership. This is a very complex issue. Should we favor, say, blacks or whites in college admission just because of ethnic origin? We believe not.

The larger problem. Jensen's book consists essentially of two parts. Up to page 366, we are given Jensen's perceptions of test theory and practice. The remainder discusses test bias. Our judgment is that the first part is a presentation of ideas that have been around for some decades. The study and theory of mental ability are, in fact, very difficult. We do not know what the natures of mental ability are. We have a statistical or data-analytic approach, which is purely descriptive, but one that is useful, of course. The second part tells us that all attempts to show predictive bias with respect to minority groups have essentially failed. One can find small differences, but they are not, we judge, important, either to the individual or to society.

There is no doubt about the one standard deviation average difference between blacks and whites with respect to most of the supposed measurements of "intelligence," nor is there doubt that similar differences occur for academic achievement, occupational status, and many other socially relevant indices. We surely have to accept the general thesis that the standards of educational development are not arbitrary; they are not chosen so as to show that one racial group is "inferior" to another. They are chosen, naturally, to reflect what the dominant population group considers to be relevant. They may be wrong in some respects, placing, perhaps, too much emphasis on certain types of ability that the majority group is led merely by prejudice to value.

This tells us that we must address the big issue, untrammeled by notions of intelligence and by general factor thinking, of why we find large differences that affect the whole of our pluralistic democratic society. In asking, Why? we have to ask, What can we do with the human of age 5 or 50, or whatever, who lacks reasonable seeing ability or muscular ability, reading ability, or numerical ability, or whatever? Suppose we have a child who has a heart murmur that is properly judged to lead to inability to live a desired type of life. We can and must ask why. But in general terms we can never answer absolutely and definitively. It may have happened because of genes, because of certain insults during parturition, or because of insults early in life, such as rheumatic fever, and so on. We can *never* nail down the causation. It is unfortunate that science in the broad does not teach this.

In the case of mental abilities we have to ask what intervention can be done to prevent or remove a paucity of ability. Our view is that the picture of mental tests given by Jensen is sterile with regard to this question, which we regard as *the* critical one. Suppose that we accept with Jensen the ideas on the general factor. Suppose we accept (as we do) that the usual tests have predictive power (and we note, parenthetically, that this experiential fact does *not* depend at all on a theory of intelligence). What are we to do? Perhaps we should not fault Jensen for the complete absence of consideration of this question. Actually, we surmise that Jensen has such high belief in an "innate" general factor that the question is not worth addressing. We do fault Jensen in this respect. We note that there is no discussion of experiments on education. We are very struck by the Heber Milwaukee experiment (Heber, Gaber, Harrington, Huffman & Falender 1972). We regard as a grave defect Jensen's perception of mental testing and its predictive validity as a purely observational-correlational activity. We see a huge need for research that leads to intervention that *causes* the inability or problem not to arise. It is becoming rather clear, we are inclined to think, that the very early environment of the child, perhaps even in the first three months, has a profound effect on the development of the mental abilities that are needed and demanded by our society. It is becoming clear to all (though it has been clear to some for many years) that adoption from a low SES origin into a family of high SES status is associated with a large increase in IQ of adoptives compared to nonadoptive children.

This research will require the full panoply of experimental design and statistical methods in general. It will require the development of testing procedures. How can one determine whether an intervention process increases "mental growth" if one cannot measure "mental ability"? How can one measure mental ability except by the development of tests? How then can anyone rationally take the view that the idea of mental testing is a weird, unscientific idea invented by a "stupid" class of people called testers or test builders? That there is in the whole area of mental testing a considerable variability of so-called expert opinion is not something to be pointed to with scorn. We had a Lamarckian genetics, we had Mendelian genetics, and we have DNA genetics. We have seen controversies at every single point in time, and the history of science tells us that if a field does not have controversies it is ossified. So when we look at Spearman, Thorndike, Burt, Jensen, and the like we see controversy. If one wishes to find backing for any particular point of view, one can find a supporting quotation from one of these leaders, and also an opposing quotation, even from the same writer. A rational judgment is that all these workers made significant contributions to the branch of investigation.

We have to recognize and accept the idea that we do not have even partial understanding of the origin and nature of mental abilities. To the parent of the 9-year-old child who has only 5-year-old reading ability and who asks, "Why is my son unable to read at the appropriate level?," we can only say, after eliminating the usual organic possibilities, "We do not know!" And then if someone asks, "Why don't you know?" we can only answer, "The minds of many workers have tried honestly with a huge mental effort to develop knowledge that would enable us to suggest strongly a particular answer, but they have been unsuccessful." And then if the remark is made, as it often is, unfortunately, that the people who work in the area are not bright, not intelligent, we can only answer, and *we should answer:* The field is open. We need good minds desperately, and we will be most happy if you can work in the field and advance our understanding."

Final remarks. We gave our summary general evaluation of Jensen's book, which was commendatory with regard to his objective, in the original BBS treatment (Kempthorne & Wolins 1980). The reader will notice that we have considerable agreement with some of the "peers" and considerable disagreement with others.

In his reply, Jensen (1980b) referred to our longer preliminary commentary and mistakes we mentioned. It is appropriate, we judge, to list now what we regard as major and pervasive defects in the exposition rather than the numerous but isolated technical mistakes in analysis. We shall be rather cryptic: (1) The exposition of the nature and role of variance and ANOVA we judge to be significantly deficient; (2) the nature and role of path analysis, similarly; (3) the nature and role of factor analysis, similarly; (4) the exposition of the role of normality of distribution is highly defective; (5) the validity of criterion measures is inadequately supported; (6) the discussion of intelligence and Intelligences A, B, and C is highly defective; (7) the discussion of achievement tests is inappropriate and insufficient; (8) the limitations of observational studies and of results of statistical analysis are not adequately appreciated (in spite of disclaimers on pages 615 and 617). We have attempted to substantiate some of these defects. With reference to (7) we note the remarks of Sternberg (1980, p. 353): "At present, at least, we have no way of measuring intelligence except through tests that, at one level or another, are achievement tests." Indeed, Jensen, himself, says (1980a, (p. 250): "No really clear distinction can be made operationally at the level of tests between *intelligence* and intellectual *achievement*." We find ourselves in strong disagreement with the review of Longstreth (1980), and we cannot accept Vandenberg's (1980) statement that Jensen "established a scientific basis for the concept of intelligence - almost an existence proof."

We close with an overall general reaction. Jensen includes brief remarks throughout that show he is aware of the problems we discuss. We shall not document this. But one example is the following: On methods of analysis of correlation into genetic and nongenetic components, he says (1980a, p. 145): "Their use with humans in natural environments would involve inordinate methodological difficulties because of correlation between genotypes and environment." What better indictment of Jensen's own heredity-IQ research can one find? - and from Jensen himself! We do not wish to make informal generalizations, but we felt that we found many examples of Jensen "riding every horse," regardless of direction.

Because the topic Jensen discusses is of such vast importance to society, his book has been discussed in many places in the literary world. In spite of the numerous criticisms we give, we have found ourselves having deep sympathy with Jensen in that reactions to the book have too often been emotional, too often based on justified criticisms of other of his writings, and too often blind to the mass of evidence offered about mental tests and their correlates. This evidence exhibits with overwhelming force that a huge social problem exists. Jensen's massive book has shown that mental tests are not biased. Jensen and the testers are *not at all* to blame for the problem. But we can blame them, perhaps, for failing to make significant suggestions for solving the problem.

ACKNOWLEDGMENT

Journal Paper No. J-10312 of the Iowa Agriculture and Home Economics Experiment Station, Ames, Iowa. Project 890.

IQ or intelligence?

Atam Vetta

Department of Mathematics, Statistics, and Computing, Oxford Polytechnic, Oxford OX3 OBP, U.K.

In his Response to my review (Vetta 1980a), Jensen (1980b) accepts some of my criticisms and ignores others. He also makes assertions concerning my views that are not correct, and he issues a challenge to me. I believe that I should not introduce new material to this discussion, but restrict myself to my review and his response.

Some research workers (for example, Thoday 1973 and Rawle 1980) had, in the past, invited Jensen to take account of my criticisms of his work. He did not respond. I am, therefore, grateful to the editors of BBS for providing a forum to enable him to reply to some of my criticisms. Naturally, I am pleased that he accepts some of them. Since his 1969 paper Jensen has accumulated a following that includes some individuals who have only a remote acquaintance with statistical and genetic concepts. It is therefore important that the full implications of Jensen's acceptance of my criticisms be stated clearly.

Jensen accepts that he was wrong in asserting that the two ratios $\bar{X}_b/\bar{X}_w = \sigma_b/\sigma_w$ differ statistically. This means that his assertion concerning the differences in the "mental growth rates" of blacks and whites is incorrect. He promises to make corrections in the next edition of his book, and this is one that should be made. He accepts my contention that his rigorous test of interval scale is meaningless. Obviously, he now understands that regression provides no proof of polygenic inheritance. He would need to make these corrections as well.

Jensen suggests that I am confusing sibling regression with the heritability issue. A careful reading of the appropriate paragraph should indicate that there are no grounds for this suggestion. No one who has studied R. A. Fisher as deeply as I have could be guilty of such a confusion. Indeed, I have shown (Vetta 1976) that assortative mating will create correlation between the additive and dominance deviations of a parent and progeny. [This was asserted by Sewell Wright (1952) without proof.] This affects the parent-child and sib correlations in a rather complex way. Therefore, the concept of narrow heritability should be used with some care when a population mates assortatively. The fact is that Jensen's work shows some confusion concerning the concept of broad heritability (see, for example, Hirsch, McGuire & Vetta 1980).

Jensen says (1980b, p. 362) that "Vetta has for a long time been a harsh critic of research on the genetics of intelligence." I am sorry; this is not quite so. I am in *favour* of research on the genetics of intelligence. However, I do not accept that an individual's score on *an* IQ test reflects his whole intelligence. I am a harsh critic of what I regard as incompetent statistical and genetic analysis of IQ data.

Concerning Jensen's challenge that I propose a different distribution of mental ability: I have indoctrinated myself too much with Fisherian genetics to contemplate any distribution other than the normal for mental ability, as distinct from IQ. This does not, however, mean that I decline his challenge, because I do know of a test constructor who doubts that mental ability is normally distributed. Jensen, of course, knows of him too. I am therefore surprised at his challenge, which he does not restrict to me alone. The name of the investigator in question is Wechsler, who said (1944) that "some authors also believe that the resulting frequency curve ought to be Gaussian [i.e. normal] or as nearly Gaussian as possible. The last requirement seems to be a result of the wide-spread but mistaken belief that mental measures distribute themselves according to the normal curve of error." He produced an IQ test that does not give a normal distribution.

If Jensen is prepared to concede that mental ability (i.e., intelligence) is not equivalent to IQ, then I shall be happy to show that the distribution of the latter in a population is not normal. D. D. Dorfman (private communication) makes an excellent point in this connection (I hope I present his view accurately): He says that if we accept the assertion that the distribution of IQ among different socioeconomic groups is normal, then it is not likely that the distribution for the whole population is normal. I hope that Jensen will take Dorfman's statement into consideration.

Elsewhere, I discuss (Vetta 1980b) the evidence produced by Jensen (1980a) concerning the normality of the distribution and conclude that it provides no grounds for asserting that the distribution of IQ in a population is normal. When that paper is published, I hope Jensen will reply.

It is true that Jensen did not cite N. D. M. Hirsch (1926), but he did cite N. D. M. Hirsch (1930). I find Jensen's statement (1980b, p. 360) that he has not "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" perplexing. Either N. D. M. Hirsch's work is good, in which case Jensen ought to respond to my question, or it is worthless, in which case one wonders why he cited him. Moreover, I was under the impression that Jensen had indeed evaluated some old studies; otherwise, why would he rely on Shuey (1966), who summarised a large number of old studies, many of which were conducted in the South.

Jensen has now taken a welcome step by accepting some of my criticisms. May I ask him to respond to my other criticisms of his work. He will find it difficult to contradict the following assertions: (1) There is no evidence to indicate that IQ is a polygenic trait and (2) even if we assume that it is, there is no way, apart from breeding experiments on human populations, to find the nature-nurture components of individual variance in IQ.

ACKNOWLEDGMENTS
I am grateful to Jerry Hirsch and his colleagues in the Institutional Racism Program for their encouragement while I was on sabbatical leave in the psychology department at Urbana-Champaign campus of the University of Illinois.

Author's Response

Bias in mental testing: A final word

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

No one reviewer can possibly be expected to critique every detail of a large and complex book at every level of analysis from main ideas to misprints. Hence the unique value of having a large number of critics: Although there is probably more chaff than grain in the sum total of multiple criticisms, there is also the probability of more potentially useful substance in the net yield. Even the least of it, such as the correction of misprints, is ultimately valuable. *Bias in Mental Testing* has been well favored in this respect, with the benefit of some 70 reviewers since its publication in January 1980, 32 of them in the BBS multiple book review [*BBS* 3(3) 1980] and this Continuing Commentary. Few works are ever subjected to such thorough and detailed scrutiny. It testifies to the importance of the book's subject matter.

The many critiques have been highly diverse and generally valuable for the advancement of research on psychometric bias. The specific problems (and, in some cases, actual errors) in some of the statistical formulae so assiduously ferreted out by **Kempthorne & Wolins** and by **Darlington & Boyce** will, of course, be checked by other qualified statisticians, for I do not assume a priori that critics are less liable to error than those they criticize. Where the technical criticisms prove valid, they will be most useful in making revisions for the second edition. For that I am indeed grateful. Although I am not a statistician, according to those I have consulted, most of the statistical issues in question (except for obvious misprints like the omission of the brackets in Fisher's r to z transformation) are far from elementary. They usually concern the estimation of standard errors for novel, often complex, indicators of item bias. These are not always routine statistics, and my proposed solutions in some cases undoubtedly fall short. Ideal formulations in these cases will depend upon the skills of mathematical statisticians like Kempthorne, and probably the statistically least tractable indices of internal bias will be discarded or replaced by more elegant and efficient methods. I doubt, however, that this will have any effect on the direction in which the preponderance of the empirical evidence so clearly points.

According to **Darlington & Boyce,** where my own formulations for estimating standard errors err, as compared with what they regard as more correct estimates, they err in the direction of detecting a given degree of bias as being statistically significant in some cases where the statistical tests suggested by Darlington & Boyce would lead to the conclusion of no bias. To suggest, as these reviewers do, however, that the statistical problems they point out in the notes of Chapter 9 are grounds for doubting the main conclusions of the book is exaggerated and untenable. **Kempthorne & Wolins,** who are statisticians and whose view of the book is much less myopic, in fact agree with all of the book's main conclusions regarding test bias. These conclusions rest on an enormous body of research by numerous investigators and my conclusions are essentially the same as those of other psychometricians who have reviewed this body of evidence. If anyone can review the total available research in this field and arrive at opposite conclusions, it will be a spectacular feat. Moreover, hardly any of the studies on which these conclusions are based have depended upon the particular statistical formulae about which Darlington & Boyce complain.

I regret that I do not find the present commentaries by **Hirsch & Tully** and by **Vetta** as helpful as the others in this collection. For one thing, they do not deal with central issues in the test bias argument, which, as I have pointed out repeatedly, should not be confused with the nature-nurture or heritability issue. Even their argumentation in this realm appears to me nihilistic and obscurantist, aimed at defending entrenched positions that are shared by few, if any, behavioral geneticists (e.g., the notion that the substantial heritability of IQ lacks evidential support). If either Hirsch or Vetta thinks he can make a coherent argument for such a position, he should do so in a full-fledged article or book. So far, their writings on this topic strike me as a hodgepodge of esoteric quibbles. (An up-to-date exposition of my stance on the inheritance of mental ability is to be found in my latest book, Jensen 1981.)

Hirsch & Tully provide no persuasive argument that the results of Harrington's (1975) experiment on different strains of rats learning mazes can either logically or empirically override conflicting conclusions on test bias based on standardized intelligence tests used in human populations. I have discussed the interpretation of Harrington's experiment in detail elsewhere (Jensen, in press). The multi-author book in which my essay appears (along with a chapter by Harrington on his experiment) can be recommended to readers who may want a more comprehensive critical discussion of test bias from diverse viewpoints than is afforded by the set of critiques in this Continuing Commentary section.

References

Bohannan, P. (1973) Heritability of intelligence. *Science* 182:115. [JH, OK]

Dorfman, D. D. (1980) Test bias: What did Yale, Harvard, Rolls Royce, and a black have in common in 1917? *Behavioral and Brain Sciences* 3:339-40. [JH, OK]

Falconer, D. S. (1960) *Introduction to quantitative genetics*. Edinburgh: Oliver and Boyd. [JH]

Gillie, O. (1980) Burt: The scandal and the cover-up. *Supplement to the Bulletin of the British Psychological Society* 33:9-16. [JH]

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

Guilford, J. P. (1956) *Fundamental statistics in psychology and education*. 3rd ed. New York: McGraw-Hill. [RBD]

Hansen, L. M.; Mendel, R. M. & Wolins, L. (1979) Three flies in the ointment: A reply to Arvey and Mossholder. *Personnel Psychology* 32:511-16. [OK]

Harman, H. H. (1967) *Modern factor analysis.* 2d ed. Chicago: University of Chicago Press. [OK]

Harrington, G. M. (1975) Intelligence tests may favour the majority groups in a population. *Nature* 258:708–9. [JH, ARJ]

(1982) An experimental model of bias in mental testing. In: *Perspectives on bias in mental testing*, ed. C. R. Reynolds and R. T. Brown. New York: Plenum Press, in press. [JH]

Heber, R.; Gaber, H.; Harrington, S.; Hoffman, C. & Falender, C. (1972) Rehabilitation of families at risk for mental retardation. Rehabilitation Research and Training Center in Mental Retardation. Madison, University of Wisconsin. [OK]

Hirsch, J. (1963) Behavior genetics and individuality understood: Behaviorism's counterfactual dogma blinded the behavioral sciences to the significance of meiosis. *Science* 142:1436–42. [JH]

(1967a) Behavior-genetic, or "experimental," analysis: The challenge of science versus the lure of technology. *American Psychologist* 22:118–30. [JH]

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

(1967c) Intellectual functioning and the dimensions of human variation. In: *Genetic diversity and human behavior*, ed. J. Spuhler. Chicago: Aldine. [JH]

(1976) IQ tests and majority groups. *Nature* 260:8. [JH]

(1981) To "unfrock the charlatans." *SAGE Race Relations Abstracts* 6:1-65. [JH]

Hirsch, J.; Beeman, M. & Tully, T. P. (1980) Compensatory education has succeeded. *Behavioral and Brain Sciences* 3:346–47. [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. S. Lockard. New York. Elsevier. [AV]

Hirsch, N. D. M. (1926) A study of natio-racial mental differences. *Genetic Psychology Monographs* 1:293–405. [AV]

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

Humphreys, L. G. (1971) Theory of intelligence. In: *Intelligence: Genetic and environmental influences*, ed. R. Cancro, pp. 31-55. New York: Grune and Stratton. [OK]

Hunt, J. McV. (1981) Review of A. Jensen *Bias in Mental Testing. BioScience* 31:151-153, 176-177. [JH]

Jensen, A. R. (1969) How much can we boost IQ and scholastic achievement? *Harvard Educational Review* 39:1-123. [JH, AV]

(1980a) *Bias in mental testing.* New York: Free Press. [RBD, JH, OK]

(1980b) Correcting the bias against mental testing: A preponderance of peer agreement. *Behavioral and Brain Sciences* 3:359-68. [JH, OK, AV]

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

(in press) Test bias: Concepts and criticisms. In: *Perspectives on bias in mental testing*, ed. C. R. Reynolds & R. T. Brown. New York: Plenum 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* 70:311–49. [JH]

Kempthorne, O. (1957) *An introduction to genetic statistics.* New York: Wiley. [JH]

(1978) Logical, epistemological and statistical aspects of nature-nurture data interpretation. *Biometrics* 34:1-23. [JH, OK]

Kempthorne, O. & Wolins, L. (1980) Controversies surrounding mental testing. *Behavioral and Brain Sciences* 3:348–49. [JH]

Lawley, D. N. & Maxwell, A. E. (1963) *Factor analysis as a statistical method.* London: Butterworths. [OK]

Longstreth, L. E. (1980) The definitive work on mental test bias. *Behavioral and Brain Sciences* 3:350–51. [JH, OK]

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]

McNemar, Q. (1949) *Psychological statistics.* New York: Wiley. [RBD]

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

Rawle, R. E. (1980) Review of A. R. Jensen, *Bias in mental testing. The Times Higher Education Supplement*, 25 April. [AV]

Reynolds, C. R. (1980) In support of *Bias in Mental Testing* and scientific inquiry. *Behavioral and Brain Sciences* 3:352. [JH]

Shuey, A. M. (1966) The testing of Negro intelligence. New York: Social Science Press. [AV]

Spearman, C. (1914) The heredity of abilities. *Eugenics Review* 6:219–37. [JH]

Sternberg, R. J. (1980) Intelligence and test bias: Art and science. *Behavioral and Brain Sciences* 3:353–54. [OK]

Thoday, J. M. (1973) Review of A. R. Jensen, 1973, *Educability and group differences. Nature* 245:418–20. [AV]

Vandenberg, S. G. (1980) An existence proof for intelligence? *Behavioral and Brain Sciences* 3:355–56. [OK]

Vetta, A. (1976) Quantitative inheritance involving assortative mating and selection. Ph.D. thesis, London University. [AV]

(1980a) Correlation, regression and biased science. *Behavioral and Brain Sciences* 3:357–58. [JH, AV]

(1980b) On the indeterminacy of nature-nurture components of IQ. Unpublished. [AV]

Wahlsten, D. (1980) Race, the heritability of IQ, and the intellectual scale of nature. *Behavioral and Brain Sciences* 3:358–59. [JH]

Wechsler, D. (1944) *Measurement of adult intelligence.* Baltimore: Williams & Wilkins. [AV]

Wright, S. (1921) Correlation and causation. *Journal of Agricultural Research* 20:557–85. [OK]

(1952) The genetics of quantitative variability. In: *Quantitative inheritance*, ed. E.C.R. Reeve & C.H. Waddington, p. 18. London: H.M.S.O. [AV]

On John R. Searle (1980) Minds, brains, and programs. BBS 3:417–457.

Abstract of the original article: This article can be viewed as an attempt to explore the consequences of two propositions. (1) Intentionality in human beings (and animals) is a product of causal features of the brain. I assume this is an empirical fact about the actual causal relations between mental processes and brains. It says simply that certain brain processes are sufficient for intentionality. (2) Instantiating a computer program is never by itself a sufficient condition of intentionality. The main argument of this paper is directed at establishing this claim. The form of the argument is to show how a human agent could instantiate the program and still not have the relevant intentionality. These two propositions have the following consequences: (3) The explanation of how the brain produces intentionality cannot be that it does it by instantiating a computer program. This is a strict logical consequence of 1 and 2. (4) Any mechanism capable of producing intentionality must have causal powers equal to those of the brain. This is meant to be a trivial consequence of 1. (5) Any attempt literally to create intentionality artificially (strong AI) could not succeed just by designing programs but would have to duplicate the causal powers of the human brain. This follows from 2 and 4.

"Could a machine think?" On the argument advanced here *only* a machine could think, and only very special kinds of machines, namely brains and machines with internal causal powers equivalent to those of brains. And that is why strong AI has little to tell us about thinking, since it is not about machines but about programs, and no program by itself is sufficient for thinking.

Stimulating understanding: Making the example fit the question

Thomas Edelson
10114 Fleming Avenue, Bethesda, Md. 20014

Searle (1980b, p. 417) tells us that his central argument is intended to show that "instantiating a computer program is never by itself a sufficient condition of intentionality." I wouldn't want to disagree with this. For something to have intentionality, it also needs to interact with the world in the right kinds of ways. If a program were set to running in an environment such that it "perceived" and "understood" stimuli coming from a simulated world, then this would at most be simulated perception and understanding. (Though I believe