DOCUMENT RESUME

ED 114 413	TM 004 906
AUTHOR	Jensen, Arthur R.
TITLE	Test Bias and Construct Validity.
PUB DATE	Sep 75
NOTE	31p.; Paper presented at the Annual Meeting of the American Psychological Association (83rd, Chicago, Illinois, September 1975)
EDRS PRICE	MF-$0.76 HC-$1.95 Plus Postage
DESCRIPTORS	Caucasians; Comparative Analysis; Culture Free Tests; Elementary School Students; Individual Differences; Intelligence; *Intelligence Tests; *Nature Nurture Controversy; Negroes; *Racial Differences; Socioeconomic Status; Standardized Tests; Statistical Analysis; *Test Bias; Test Reliability; *Test Validity
IDENTIFIERS	Peabody Picture Vocabulary Test; Ravens Progressive Matrices; Stanford Binet Intelligence Scale; Wechsler Intelligence Scale for Children; Wonderlic Personnel Test

ABSTRACT

The several statistical methods described for detecting test bias in terms of various internal features of a person's test performances and the test's construct validity can be applied to any groups in the population. But the evidence regarding groups other than U.S. blacks and whites is either lacking or is still too sketchy to permit any strong conclusions. The evidence regarding black-white comparisons, however, is based on a number of well-known, widely used, and quite diverse standardized individual and group tests of intelligence given to a large representative sample of whites and blacks. The results are unequivocal: none of the several subjective indices of cultural bias shows any significant indication of bias in any of these tests when they are used with blacks and whites. Correlation of raw scores with age, internal consistency reliability, rank order of item difficulty, relative difficulty of adjacent items, item correlation with total score, loadings of items or tests on the general factor, and relative frequencies in choice of error distractors--all are substantially the same in black and white groups. It is concluded that these standardized tests of intelligence--the Peabody Picture Vocabulary, Raven's Progressive Matrices, Wechsler Intelligence Scale for Children, Stanford-Binet, Wonderlic Personnel Test, and most likely other similar tests--are not at all culturally biased for blacks and whites. They behave statistically the same in both racial groups and do essentially the same job in both groups. (Author/DEP)

TEST BIAS AND CONSTRUCT VALIDITY

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

Invited Address presented at the 83rd Annual Convention of the American Psychological Association, in Chicago, Illinois, on September 2, 1975.

Test Bias and Construct Validity

Arthur R. Jensen

University of California, Berkeley

Most psychologists are surely familiar with the claims of critics that our mental tests are culturally biased against certain minorities, especially blacks, and are culturally biased in favor of middle class whites. As a reminder, here are just a few direct quotations I have picked up from the literature. They are all very typical.

> "IQ tests are Anglocentric; they measure the extent to which an individual's background is similar to that of the modal cultural configuration of American society."

> "IQ measures everyone by an Anglo yardstick. There is a conspiracy to make a narrow, biased collection of items the real measure of all persons."

> "Persons from backgrounds other than the culture in which the test was developed will always be penalized."

> "Intelligence tests are sadly misnamed because they were never intended to measure intelligence and might have been more aptly called CB (cultural background) tests."

> "IQ tests yield the best results when taken by those who come from the same cultural background as the devisers of the tests."

> "Tests are clearly discriminatory against those who have not been exposed to the culture, entrance to which is guarded by the tests."

"Racial, ethnic, and social class differences in mean IQ scores may not be due to genes or environment, but are probably inherent in the psycholinguistic, cultural, and temporal biases of the test."

"There are enormous social class differences in a child's access to the experiences necessary to acquire the valued intellectual skills."

"Aptitude tests reward white and middle class values and skills, especially ability to speak Standard English, and thus penalize minority children because of their backgrounds."

"The middle-class environment is the birthright for IQ test-taking ability."

"The IQ test is a seriously biased instrument that almost guarantees that middle-class white children will obtain higher scores than any other group of children. The more similar the experiences of two people, the more similar their scores should be."

"IQ scores reported for blacks and low socioeconomic groups in the U.S. reflect characteristics of the test rather than of the test takers."

"Culturally unfair tests may be valid predictors of culturally unfair but nevertheless highly important criteria. Educational attainment, to the degree that it reflects social inequities rather than intrinsic merit, might be considered culturally unfair."

"The poor performance of Negro children on conventional tests is due to the biased content of the tests, that is, the test material is drawn from outside the black culture."

"The words included in vocabulary tests are based on the frequency of their usage by whites. Blacks, who have differing vocabularies, may do poorly."

Notice the main themes in these criticisms of mental tests:

1. The tests draw heavily upon specific middle-class cultural knowledge and linguistic usage.
2. The implication is that blacks or other minorities in the U.S do not share a common culture or background of verbal and cognitive experience which is sampled by the tests.
3. Similarity in test performance is a direct function of similarity in cultural background.
4. The biggest differences in IQ scores are between lower and middle social classes and majority and minority racial groups.
5. Culturally biased tests may nevertheless show good predictive validity for predicting culturally biased criteria, like educational attainment and success in certain occupations.

Where Do IQ Tests Show Differences?

First of all, let's gain a bit of perspective as to just where tests show differences and how big those differences are relative to one another. I have been able to do this with a number of different intelligence tests, using very large samples of school children in California. I'll use the Wechsler Intelligence Scale for Children-Revised (WISC-R), as an example, with data on Full Scale IQs of more than 600 whites and 600 blacks representing a random sample of California school children, ages 5 to 12.[1]

Table 1 shows an analysis of variance, with the percentage of total variance attributable to each of the sources. The figures easiest to grasp

- - - - - - - - - - - - - - - - -

Insert Table 1 about here

- - - - - - - - - - - - - - - - -

are those in the last column, giving the average absolute difference in IQ. We had a 10-point scale of socioeconomic class on these children. The average IQ differences between all possible comparisons of the 10 social classes (within each racial group) was only 6 IQ points. (The largest SES difference was 26 IQ points in the whites and 12 IQ points in the blacks.)

The average race difference, independently of socioeconomic status (as measured by Duncan's SES index) is 12 IQ points. But here is the important point: the average difference between full siblings within the same family is also 12 IQ points. If the Wechsler IQ test is so culturally biased, as come critics claim, what kind of bias is it that produces as large a difference between siblings as between blacks and whites? Or a larger difference than the average difference between social classes? Notice, too, that the average IQ difference between families within the same social class (on a 10-point scale of SES) is 9 points, which is 33% greater than the average difference between social classes.

In short, the notion that IQ tests discriminate the most between races or social classes is just a myth. The IQ shows as much or more difference among children in the same family, sharing the same parents and culture and linguistic background, as between racial or social class groups. The generalization is just not true that the more alike is the background of two individuals,

Table 1

Estimated Percent of Variance and Average Absolute Difference in WISC-R IQ Independently Associated with Race (White-Black), Social Class, and Between and Within Families

Source	% Variance		Average IQ Difference
Social Class (Within Races)	8	22	6
Race (Within Social Classes)	14		12
Between Families (Within Race and Social Class)	29	73	9
Within Families (Siblings)	44		12
Measurement ~~Error~~	5		4
Total Sample	100		17

Sample size: Whites = 622; blacks = 622.

the more alike will be their scores on a standard IQ test. That is true only when the two individuals are identical twins.

Criteria of Cultural Bias

First, we must clearly distinguish between two concepts: *culture loading* and *culture bias*. *Culture loaded* does not mean the same as *culture biased*. Tests and test items can be ordered along a continuum of culture loading, which is the specificity or generality of the informational content of the test items. The narrower or less general the culture in which the test's information content could be acquired, the more culture loaded it is. A test may contain information that could only be acquired within a particular culture. This can usually be determined simply by examination of the test items. The specificity or generality of the content corresponds to its cultural loading. The question "Name three parks in New York City" is, in this sense, more culture-loaded than the question "How many 10¢ postage stamps can you buy for $1?"

Whether the particular cultural content causes the test to be biased with respect to the performance of any two (or more) groups in the population is a separate issue. To the extent that the test contains cultural content that is generally peculiar to the members of one group but not to the members of another group, it is liable to be biased with respect to comparisons of the test scores between the groups or predictions based on their scores.

Score differences per se, whether between individuals, social classes, or racial groups, obviously cannot be a proper criterion of bias. There is no basis for assuming *a priori* that any two populations should be equal in whatever it is that the test is supposed to measure.

Legitimate criteria of test bias are of two general types: external and internal, or predictive validity and construct validity.

For practical uses of tests, predictive validity is crucial. One criterion of test bias is if the intercepts and slopes of the regression of criterion measures on test scores differ appreciably for the two populations in question. In other words, the test scores do not predict equally well for both groups. The person's predicted performance on the criterion--job, school, etc.--will be influenced by his group membership and not just his test score. An unbiased test, on the other hand, is colorblind. It makes the same prediction of your future performance based just on your test score and the prediction turns out just as accurately whether you are white or black.

Reviews of the research on this point comparing white and black samples are unequivocal with respect to the prediction of scholastic and job performance by means of standard tests. There is a negligible difference in the slopes and intercepts of regression lines for whites and blacks. A single regression equation predicts equally well for both racial groups (Humphreys, 1973; Linn, 1973). Interestingly, the few exceptions reported in the literature would favor the black groups if the tests were used for selection, i.e., the difference in the regression lines is such that for any given test score whites slightly out-perform blacks on the criterion. In brief, the overwhelming evidence on the predictive validity of standard tests indicates that they are not biased against blacks when compared with whites. (There are too few studies of other ethnic groups to permit any general conclusions about them.)

Construct Validity criteria of test bias are more complicated, but no less important. It is very likely that tests which show little or no bias in terms of the indices of construct validity are also unbiased in predictive validity.

Construct validity criteria of bias refer to internal characteristics of the test and the degree of similarity of their statistical properties from one group to another. Construct validity, in the context of test bias, also involves the question of whether a test, or a battery of tests, measures individual differences in the same hypothetical ability in both of the populations in question. Does our theory of what the test measures yield predictions that are empirically borne out in the one group as well as in the other? If there is a difference in group means on the test, does our theory of what the test measures predict other previously unsuspected differences between the two groups?

I shall illustrate the application of some of the criteria of internal or construct bias on a variety of well-known standard tests of mental abilities, mainly intelligence or IQ tests. In all the examples, the populations for which evidence of test bias was sought by these criteria are whites and blacks in the United States. We have more extensive test data on these two groups than on any others in our population, and controversy over test bias has revolved largely around the well-known white-black differences in test scores.

Tests at the Extremes of Culture-Loading

First, let us contrast two tests that I believe most psychologists will agree are widely separated on the culture-loading continuum--the Peabody Picture Vocabulary Test (PPVT) and Raven's Progressive Matrices.

The PPVT consists of 150 plates, each with four pictures. The examiner names one of the pictures and the subject is asked to point to it. The vocabulary ranges from very easy, common, and concrete words to very rare words and abstract concepts. The Progressive Matrices consists of 60 plates, each with a missing part which the subject must select from a multiple-choice set of six to correctly complete the pattern. Items range in complexity and difficulty from a level that is passable by most three-year-olds up to a level of difficulty beyond the capacity of the average adult. Figure 1 shows typical PPVT and Raven items of moderate difficulty.

- - - - - - - - - - - - - - -

Insert Figure 1 about here

- - - - - - - - - - - - - - -

Both of these tests were individually administered to about 600 white and 400 black children, ages 6 to 12, in California schools. (Full details of this study are given by Jensen, 1974). The two groups show the typical IQ difference of about one standard deviation (15 points) on both tests.

Correlation of Raw Scores with Age. The first indication that the Peabody and Raven behave quite similarly in both racial groups is the fact that the groups are about the same in the correlation between raw scores and age in months, a correlation of about 0.70, for both tests in both racial groups. If the tests were measuring something quite different in both groups, it seems unlikely that the scores would have nearly the same correlation with age in each group.

Internal Consistency Reliability. The internal consistency reliability coefficient in the Peabody is .96, both for whites and for blacks; the Raven reliabilities for whites and blacks are .90 and .86. (The Raven has a lower reliability than the Peabody only because the Raven consists of fewer items. Corrected for length of test, the Raven's reliability is higher than the Peabody's.)

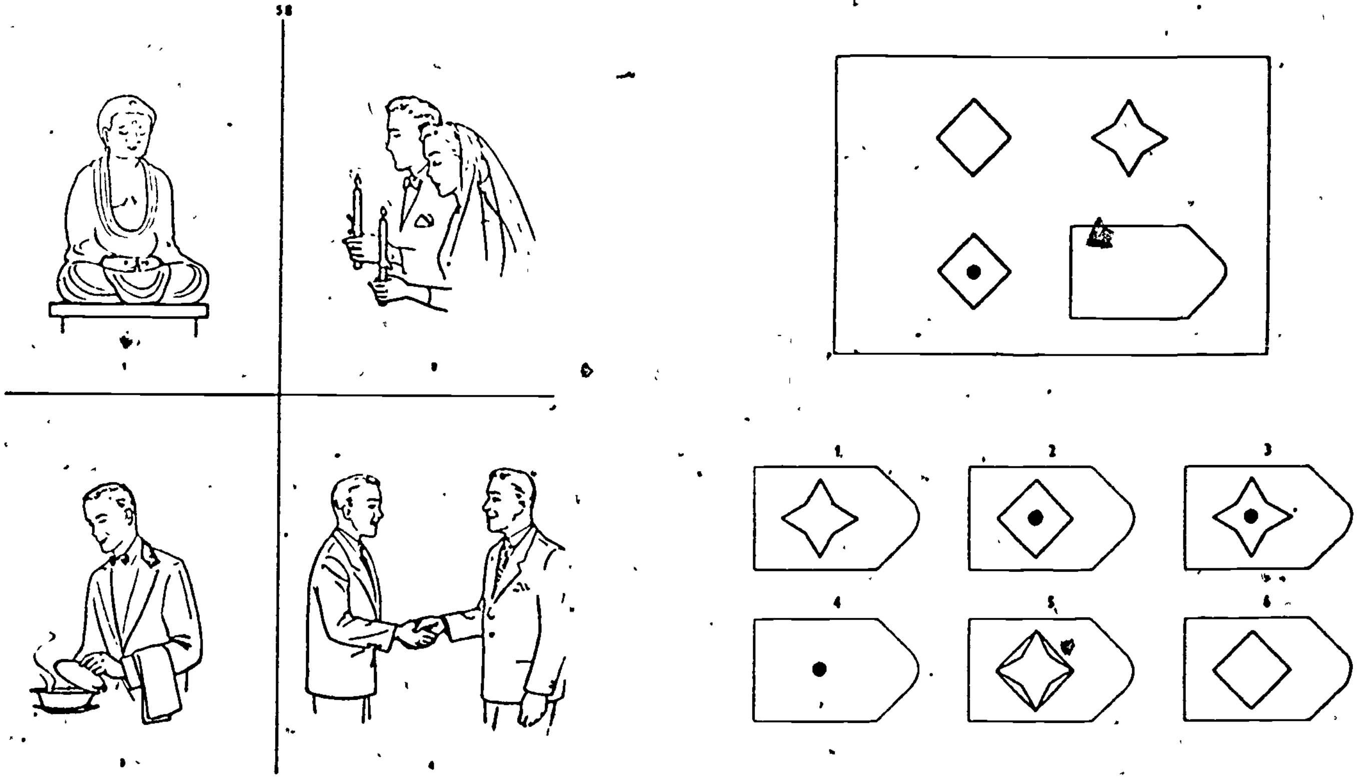

Fig. 1. Sample items of the Peabody Picture Vocabulary Test (left) and Raven's Progressive Matrices. The PPVT word for this item is "ceremony."

If one group were more careless than the other in taking the tests, or made more haphazard guesses at the answers, or otherwise contaminated their performance, we should expect quite different internal consistency reliabilities. But we see that the reliabilities are highly comparable for whites and blacks.

<u>Rank Order of Item Difficulty</u>. The percentage *P* of the group passing an item is an index of item difficulty. We can compare the rank order of these *P* values in the white and black groups and express the degree of similarity between the groups by means of the correlation between the *P* values. (All the correlations are corrected for attenuation, using the correlation of each racial group with itself, i.e., the reliability of the rank order of *P*s within each racial group.)

On the Peabody test, the correlation between rank order of item difficulty for blacks and whites is .987. The correlation between black males and black females is .983. In other words, the rank order of item difficulties on the Peabody is not as different between whites and blacks as between black males and black females. (The correlation between white males and females is .988.)

The cross-racial correlations of item difficulties in the Raven are all .99 or greater when corrected for attenuation.

We can safely conclude that for the Peabody and the Raven, the rank order of item difficulty is the same for whites and blacks.

This was found not to be the case when Peabody tests were obtained on white school children in London, England, as compared with age-matched white children in California. Quite a number of items differed markedly in rank order of difficulty, and some were as many as 50 items apart in rank order for Londoners and Californians. Obviously the linguistic backgrounds of

Londoners and Californians differ very much more than of whites and blacks residing in California. The English children, however, also found certain words much easier, while some were more difficult, so that the overall differences average out and both the English and the California white children obtain about the same mean IQ. California blacks, however, have a lower percent passing on every item in the test, but the rank order of item difficulty for the blacks is the same as for whites.

If the Peabody Picture Vocabulary Test were really reflecting a cultural background difference between whites and blacks, we should expect to see the kind of differences in rank order of difficulty that we see between Londoners and Californians. But we find no difference between blacks and whites in the rank order of item difficulties.

Correlation of P Decrements. Let's remove the level of item difficulty altogether and look at only the *differences* between item difficulties for adjacent items in the test. This is P_1-P_2, P_2-P_3, and so on, where P_1 is the percent passing item 1, P_2 is the percent passing item 2, and so on. This is a most sensitive index of group similarity. On this index, called the *P* decrement, the equivalent Forms A and B of the Peabody test are correlated zero in the very same group of persons, even though the correlation of item difficulties for Forms A and B in the same group is .97.

The correlation (corrected for attenuation) between whites' and blacks' *P* decrements on adjacent items is .830. The correlation between *P* decrements of males and females is .823 in whites and .880 in blacks. Thus, we see again that the two races differ no more than do the two sexes of the same race.

The Raven's *P* decrements in whites and blacks correlate .980.

If the items of these tests were culturally biased for blacks, it would be remarkable indeed that their rank order of difficulty and the differences in difficulty between adjacent items should be virtually the same in both the black and white groups: It would seem more remarkable that two tests as dissimilar in culture-loading and information content as the Peabody and the Raven should both show such high degrees of similarity between blacks and whites in the rank order of P values and P decrements.

Matching Peabody and Rave: Items. Are verbal tests more biased than nonverbal? The small differenc s between the P abody and Raven that we have seen in the preceding analyses show very little difference between the tests on the two indices of bias we have examined.

Going a step further, we perfectly matched P abody and Raven items for difficulty in the white group. For each of 35 Raven items we found a Peabody item with exactly the same percent passing. If the culture-loaded Peabody was more biased against blacks than the Raven, then we should expect blacks to obtain lower scores on the Peabody than on the Raven, when the difficulties of the two tests are perfectly matched in the white group. It turned out that blacks showed no significant difference between Raven and Peabody scores. Raven and Peabody items matched for difficulty in the white group, it turns out, are thereby also matched for difficulty in the black group.

We tried the same analysis on a Mexican-American group. But it showed a highly significant difference in favor of the Raven. Thus there is some evidence that a vocabulary test in English may be a biased test of intelligence for Mexican-Americans.

For reasons I need not go into here, I don't think the Peabody is an especially good measure of general intelligence for either whites or blacks. But I find no evidence that it is biased with respect to either of these groups.

Item Discriminabilities Within and Between Racial Groups

In both the Peabody and the Raven we compared (a) the correlations between single items and total score within each racial group , and (b) the point-biserial correlations between single items and the racial dichotomy. The first set of correlations, a, tells us how well each item measures whatever the test as a whole is measuring and how well the item discriminates among persons within a given racial group. The second set of correlations, b, tells us how much the items discriminate between the two racial groups. It turns out that the items that best measure individual differences within each racial group are the very same items that discriminate the most between the racial groups. These items have the highest correlations with total score for both blacks and whites.

Analysis of Wrong Answers

Culture bias leads to the expectation that whites and blacks should make different errors among the multiple-choice distractors of the items they get wrong. But analysis of incorrect responses (errors) in the Peabody shows that the errors are distributed in a non-chance fashion over the multiple-choice distractors for each item in the same proportions for whites and blacks. There were several significant exceptions to this finding, in Raven's Matrices: on some items blacks made different errors than whites. But in every such instance it was found that the black children's proportions of responses to the various error distractors were the same as the proportions for white children who were approximately two years younger in chronological age. Thus it appears that the few differences that were found between white and black children are more clearly related to differences in level of mental maturity than to cultural differences.

Simulation of White-Black Differences

An overall analysis of variance was performed on the following factors and all their interactions, for both the Peabody Picture Vocabulary and Raven's Matrices: Race, Sex, Age, Items, and Subjects.

The interaction of greatest interest in terms of detecting culture bias is the Race × Items interaction. The size of the Race × Items interaction, relative to other sources of variance, is a sensitive index of bias. It turns out that the interaction, though statistically significant, accounts for less than 1 percent of the total variance in both the Peabody and the Raven.

We found that we could perfectly simulate, within the margin of sampling error, this whole analysis of variance, with all its main effects and all their interactions, using only the white sample. We called this comparison of two different age groups of whites a Pseudo-race comparison.

We divided the entire white sample into two groups: a younger group (ages 6 to 9), and a slightly overlapping older group (ages 8 to 11). The same analysis of variance that was performed on blacks and whites when performed on these two different age groups of whites reproduced all of the features of the analysis of variance on the two racial groups. There is just no difference between the two sets of variances, within the margin of sampling error. This is true for both the Peabody and the Raven. The Pseudo-race × Items interaction was also about 1 percent of the variance.

Finally, by doing the same analysis again on the two races, but this time using whites of ages 6 to 9 and blacks of ages 8 to 11, we found that the Race × Items interaction became quite nonsignificant (less than 0.2 percent of the total variance).

Further analyses in this vein failed to reveal any features of the Peabody or Raven performance which will statistically distinguish blacks from whites who are about two years younger, or which show any differences between blacks and whites (of the same age) that do not show up also between groups of younger and older whites.

In the light of these findings, for anyone to maintain that these tests are culturally biased with respect to black-white comparisons, he would have to argue that the cultural differences between California blacks and whites perfectly simulate age differences within the white group, for such a diversity of indices as rank order of item difficulties, P decrements, interitem correlations, choice of distractors, and item factor-loadings on the first principal component--on tests as diverse as picture vocabulary and progressive matrices! Obviously such an argument is grossly implausible.

A variety of other tests have shown the same sort of thing; that is, black-white differences in test performance can be perfectly simulated, quantitatively and qualitatively, by comparing groups of younger and older white children. This has been shown for Piagetian conservation tests, copying simple geometric designs, and developmental tests involving free-choice preferences for matching stimuli on the basis of color, form, size, and number (Jensen, 1975).

Indices of Internal Bias Applied to Other Tests

The types of analysis described above have been applied to other tests as well, all with highly similar results. But certain outstanding points are worth mentioning.

Stanford-Binet. The rank order of difficulty correlated between racial or cultural groups gains greater cogency when the test items are more heterogeneous, since it is so unlikely that a cultural difference between two groups would result in the same rank order of difficulty in the two groups over a set of items that differ markedly in their specific demands on knowledge and skills.

There is probably no more heterogeneous collection of intelligence test items to be found anywhere than the Stanford-Binet items included in the tests for ages 3-1/2 to 5. The items involve size comparisons, simple picture puzzles, discrimination of animal pictures, sorting colored buttons, verbal comprehension, picture vocabulary, opposite analogies, aesthetic comparisons, following directions, and so on.

In a doctoral thesis, Paul Nichols (1972) analyzed 16 items of the Stanford-Binet from year III-6 through IV-6--the most heterogeneous sequence of items in the whole test--given to 2,514 black and 2,526 white children, all between 4 and 5 years of age.

Note three important points: we are dealing with only a restricted portion of the Stanford-Binet test (16 items from year III-6 through IV-6), all the children are within a one-year age interval, and all are preschoolers--they haven't yet been exposed to the common culture of public schooling.

The correlation between the blacks and whites in the percent passing each of these 16 Stanford-Binet items turns out to be .96. That's .96, without correction for attenuation.

The P decrements correlate across races .50, which indicates considerable racial similarity even in the differences in difficulty between adjacent items.

Thus, in this age range, at least, the Stanford-Binet IQ test doesn't look at all culture biased, I would be quite surprised if black-white comparisons turned out very differently from this for any other section of the Stanford-Binet for any other age range.

It can also be noted that those items that critics most often single out as examples of racially biased items either have the same rank order of difficulty for blacks as for whites or are relatively easier items for the blacks, which is just the opposite of the popular claims of culture bias against blacks.

Wechsler Intelligence Scale for Children. The WISC provides some striking examples of how invalid are the critics' subjective armchair analyses of cultural bias in specific test items. For example, a favorite target of test critics is the WISC Verbal Comprehension item: "What is the thing to do if a fellow (girl) much smaller than yourself starts to fight with you?" This item is often claimed to be culturally biased against blacks, and even Dr. David Wechsler himself conceded to this claim in an interview with Dan Rather on the recent CBS-TV program "The IQ Myth."

After seeing the CBS "Myth" program, a psychology graduate student, Frank Miele, had the innovative idea of looking up the item statistics on this and other WISC items. He obtained WISC tests on large samples of age-matched white and black school children in Georgia and looked at the rank order of difficulty of this purportedly biased item within each racial group. When the easiest item in the whole WISC is ranked 1 and the hardest is ranked 161, the rank order in difficulty of the "pick a fight" item is only 42, within the black group, as compared to 47 within the white group. In short, this particular item is relatively eaiser for blacks than for whites! The armchair claims of bias are thus easily debunked by just looking at the item statistics.

The cross-racial correlation for rank order of difficulty over all 161 of the WISC items is .95. The correlation across the sexes within each racial group is .97. The correlation of difficulty rank in whites with that in blacks who average two years older is .96. Note that the WISC items, much like the Stanford-Binet items, are also very heterogeneous. Yet the rank order of difficulty of WISC items is not significantly different for whites and blacks.

Wonderlic Personnel Test.--This is a widely used general intelligence test for adults, made up of 50 very heterogeneous items--verbal, nonverbal, spatial, numerical, logical, and so on. We have found that the correlation in percent passing the 50 items, between samples of more than 700 blacks and 700 whites, is .94. The P decrements correlate .81.

We also tried to find out if 5 black and 5 white psychologists could sort out the 8 most and the 8 least racially discriminating items when all 16 items were presented on separate cards randomly shuffled. The judges sorted no better than chance. Again, armchair inspection of items is shown to be a very poor clue as to which items will discriminate the most or the least between blacks and whites.

On the other hand, we found that if you factor analyze all the item intercorrelations within each racial group, the item's loading on the general factor (or first principal component) correlates substantially with the item's racial discriminability, and this is true within both racial groups. In other words, the more highly a test item is correlated with the most general factor common to all the items, within either racial group, the more highly does the item discriminate between the racial groups.

Is g the Same g in Blacks and Whites?

The general intelligence factor or g can be defined as the first principal component--the largest single source of individual differences--in a heterogeneous collection of cognitive tests. An important criterion of the construct validity of any test (or test item) as a measure of intelligence is its loading on g when it is factor analyzed among a battery of other tests, preferably tests that are heterogeneous in informational content and in the types of cognitive processes involved in arriving at the correct answers.

How similar is this general factor for blacks and whites given the same battery of cognitive tests?

Frank Miele and R. T Osborne (personal communication) have sent me correlational data on 541 white and 237 black children in Georgia schools. All the children were given 29 cognitive tests of the greatest variety--verbal, numerical, spatial, nonverbal reasoning, form board, vocabulary, arithmetic, spelling--you name it. The tests were borrowed from several different standard batteries.

A principal components analysis was done separately in the white and black samples. Also, each racial group was randomly split in half and a principal components analysis was done in each of the split-half subgroups. In this way we can determine the reliability of the first principal component or _g_ factor _within each_ racial group.

The final step was to determine the correlation between the _g_ factor loadings, one set based on blacks and one set based on whites, over the 29 tests. This correlation turned out to be .68. Corrected for unreliability, using the within-race split-half correlations in the usual correction-for-attenuation formula, the corrected correlation becomes .97. This high correlation constitutes very strong evidence that the _g_ factor in this large battery of tests is the same _g_ for blacks as for whites.

Nichols (1972) intercorrelated 7 of the subtests of the Wechsler Intelligence Scale for Children (WISC) combined with the Bender-Gestalt Test, the Draw-a-Man Test, the Illinois Test of Psycholinguistic Abilities, and tests of reading, spelling, and arithmetic achievement--13 tests in all. This test battery was factor-analyzed separately in a group of 986 whites and 975 blacks, all 7 years of age, drawn from Boston, Philadelphia, and Baltimore. The _g_ loadings of the 13 tests correlate .98 across the races. (That's .98 without correction for attenuation.)

I have done the same cross-racial correlation of g-loadings on a battery of 14 diverse cognitive and achievement tests in large samples of blacks and whites in Grades 5 through 8. The cross-racial correlations of g loadings are of about the same magnitude as the correlation of each racial group with itself from one school grade to the next. Corrected for attenuation, the cross-racial g correlations fluctuate around unity.

I have not found any evidence based on substantial or representative groups of blacks and whites that the g factor measured by our standard tests is in the least a different g in blacks than in whites.

If the tests were culturally biased for these two populations, we would hardly expect the magnitude of the bias to be so uniform over all types of items and tests that they would all have the same g loadings (within the margin of sampling error) in black and white populations.

What is the Nature of g?

What is this g factor that practically all cognitive tests have in common despite the great diversity of their content and the seemingly different mental processes they call upon? No one really knows yet what makes for g, certainly not in any basic physiological sense. But we do have some idea as to its psychological nature.

By inspecting the g loadings of dozens of tests and many hundreds of individual items, I am led to the conclusion that the key word regarding g is complexity--complexity of the mental operations required by a test item in order for the person to produce the correct answer. Not difficulty per se, but complexity is the key to g. Items that require some active mental manipulation, some conscious mental transformation of the input, rather than just sensorimotor and short-term memory ability or a habitual response, are the most _g_-loaded items. The more mental manipulation and transformation an item involves, the more it is _g_-loaded. This is true for blacks and whites alike. I daresay it's true for all humans, and perhaps even for all animals that possess a cerebral cortex.

If we hypothesize that the well-established average IQ difference of about 15 points between blacks and whites is mainly a difference in _g_, in the sense of a capacity for dealing with cognitive complexity in any form, rather than as just a difference due to specific cultural content in the IQ test, then we should predict that blacks and whites will differ less in performance on tasks involving lesser cognitive complexity than on tasks involving greater cognitive complexity. What do we find?

Reaction Time Studies. One experimental test of this complexity hypothesis is based on differences in simple and choice reaction time to visual and auditory stimuli. In all persons, reaction time (RT) increases as a function of stimulus complexity, i.e., the number of bits of information in the signal to which the person responds. It has also been shown that there is no correlation between simple RT and IQ, but there is a negative correlation between IQ and _choice_ RT. That is, persons with higher IQs show quicker RT in a choice situation.

Four independent experiments using quite different methods but comparing simple and choice RTs in whites and blacks all show no significant race difference for simple RT. But they all show a significant race (or race confounded with SES) difference for choice of complex RT (Bosco, 1970, Jensen, 1975, Noble, 1969; Poortinga, 1972). In these experiments, each person acts as his own control. It is the _difference_ between simple and choice RT that is of primary interest, not their absolute values. Blacks, on the average, show a larger difference between simple and choice RT than do whites. RT, incidentally, is measured independently of total movement time, which is only slightly correlated with RT and is unrelated to complexity. It should be remembered that a 2-choice, 4-choice, or 8-choice RT task is still a very low level of complexity as compared with most IQ test items, but it is still more complex than the practically zero complexity of simple RT.

Forward and Backward Digit Span Memory. If g reflects capacity for mental manipulation and transformation, and if it is the g factor on which blacks and whites essentially differ, then we should expect a larger racial difference on those tests requiring more mental manipulation and transformation of the input in order to arrive at the output.

The forward and backward digit span tests of the Wechsler (WISC) lend themselves nicely to a test of this hypothesis. For one thing, most clinical psychologists judge the digit span test to be one of the least culture-loaded subtests in the Wechsler battery. Moreover, digit span shows the smallest average white-black difference of any of the subtests.

Everyone, I think, would agree that backward digit span--repeating a series of numbers in reverse order--calls for somewhat more mental manipulation and transformation than does forward digit span.

This being so, our theory of g should predict the following:

1. Backward digit span should correlate more highly with total IQ than should forward digit span.
2. Blacks and whites should differ more on backward than on forward digit span.

We tested these predictions in age-matched samples of 622 blacks and 622 whites randomly drawn from California schools[1/] (Jensen & Figueroa, in press).

Both predictions are fully borne out by the data. We found that backward span correlates significantly higher with total IQ than does forward span; and this is true <u>within</u> each racial group. We also found that the difference between whites and blacks in <u>backward</u> memory span is more than twice as large as the difference in <u>forward</u> memory span. When we control for socioeconomic status, there is no significant race difference in forward memory span, but the race difference remains substantial in backward memory span.

Figure 2 shows the total WISC IQs as a function of race and Duncan's index of socioeconomic status.

Insert Figure 2 here

Figure 3 shows forward and backward digit span scores as a function of race and SES. (The interaction of race × forward vs. backward span is significant beyond the .001 level.)

Insert Figure 3 here

Thus, the theory of *g* as a capacity for dealing with complexity and the conscious transformation of input has predicted two previously unknown phenomena: (1) the differential correlation of forward and backward digit span with IQ, and (2) the significantly smaller racial difference in forward than in backward digit span. I don't know of any hypothesis invoking cultural bias in the Wechsler tests that would have predicted either of these interesting psychological phenomena.

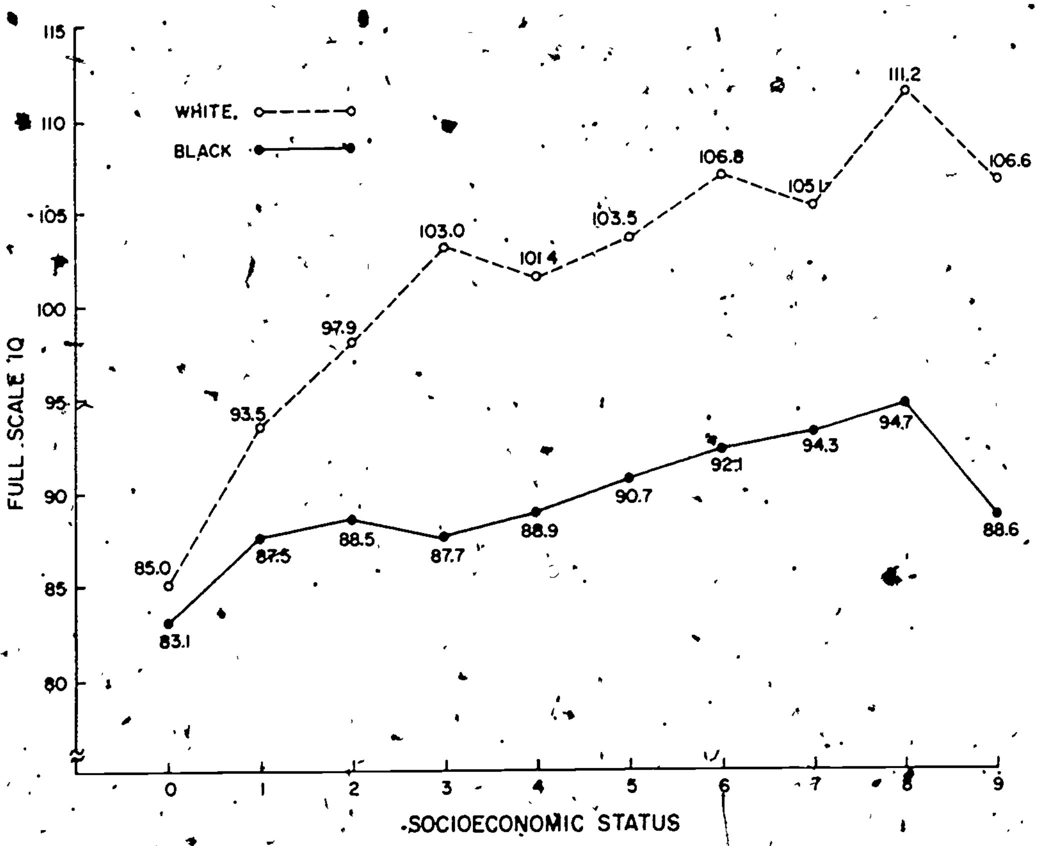

Fig. 2. WISC-R Full Scale IQ of Black (*N* = 622) and White (*N* = 622) samples as a function of socioeconomic status as measured on Duncan's Index of SES.

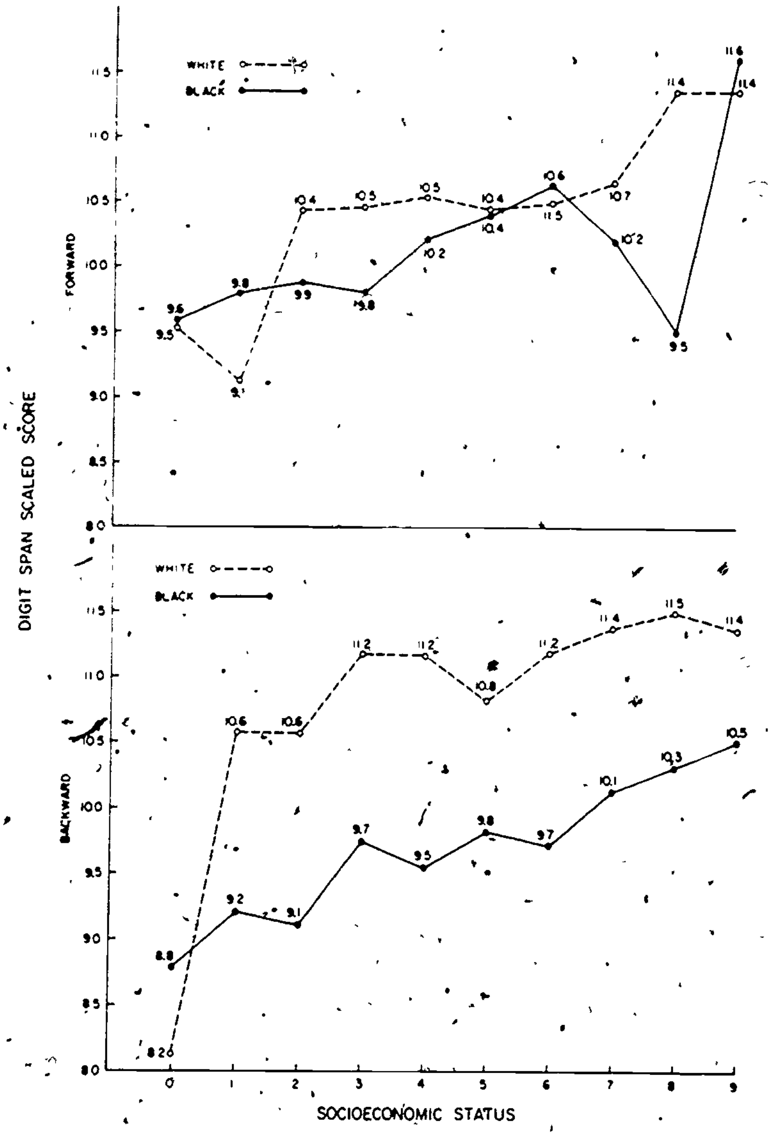

Fig. 3. WISC-R Forward and Backward Digit Span scaled scores ($\overline{X}$ = 10, σ = 3) of Black and White samples as a function of socioeconomic status.

Conclusion

The several statistical methods I have described for detecting test bias in terms of various internal features of persons' test performances and the test's construct validity can of course be applied to any other groups in the population. But the evidence regarding groups other than U.S. blacks and whites is either lacking or is still too sketchy to permit any strong conclusions.

The evidence regarding black-white comparisons, however, is based on a number of well-known, widely used, and quite diverse standardized individual and group tests of intelligence given to large representative samples of whites and blacks.

The results are unequivocal: none of the several objective indices of cultural bias shows any significant indication of bias in any of these tests when they are used with blacks and whites. Correlation of raw scores with age, internal consistency reliability, rank order of item difficulty, (i.e., percent passing), relative difficulty of adjacent items, item correlation with total score, loadings of items or tests on the general factor, and relative frequencies in choice of error distractors--all are substantially the same in the white and black groups.

I conclude that these standardized tests of intelligence--the Peabody Picture Vocabulary, Raven's Progressive Matrices, Stanford-Binet, Wechsler Intelligence Scale for Children, Wonderlic Personnel Test, and most likely many other similar tests--are not at all culturally biased for blacks and whites. They behave statistically the same in both racial groups and do essentially the same job in both groups.

Claims based on subjective armchair surmise and speculation about cultural biases in specific test items--the sole method of those critics of tests who wish to foster the myth of culture bias--are proven false by the objective evidence. Moreover, the fact that it may be possible to specially devise culturally biased items in no way proves that all of our existing standard tests are culturally biased. Culturally loaded--of course. But not culturally biased. The distinction is crucial. The myth of culture bias thrives on obscuring this distinction.

The large general factor measured by our standard tests of intelligence is clearly the same factor in blacks as in whites. The hypothesis that this general factor is a capacity for cognitive complexity, conscious mental manipulation and transformation of stimulus inputs, has led to predictions that are borne out empirically at a high level of significance.

Neither science nor the cause of social justice is served by denying these findings. As researchers our response is to question, analytically criticize, replicate results, determine their limits as to other mental tests and populations, seek the causes of test score variance, pit alternative theories against one another--and openly renounce those hypotheses that objective evidence repeatedly disproves.

References

Bosco, J. J. Social class and the processing of visual information. Final Report, Project No. 9-E-041, Contract No. OEG-5-9-325041-0034(010). Office of Education, U. S. Dept of Health, Education, and Welfare, May 1970.

Humphreys, L. G. Implications of group differences for test interpretation. Assessment in a Pluralistic Society. Proceedings of the 1972 Invitational Conference on Testing Problems. Princeton, N. J.: Educational Testing Service, 1973. Pp. 56-71,

Jensen, A R. How biased are culture-loaded tests? Genetic Psychology Monographs, 1974, 40, 185-244.

Jensen, A. R. Race and mental ability. In J. F. Ebling (Ed.), Racial Variation in Man. New York: Academic Press, 1975.

Jensen, A. R., & Figueroa, R. A. Forward and backward digit span interaction with race and IQ. Journal of Educational Psychology, in press.

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

Nichols, P. L. The effects of heredity and environment on intelligence test performance in 4 and 7 year old White and Negro sibling pairs. Unpublished doctoral dissertation, University of Minnesota, 1972.

Noble, C. E. Race, reality, and experimental psychology. Perspectives in Biology and Medicine, 1969, 13, 10-30.

Poortinga, Y. A comparison of African and European students in simple auditory and visual tasks. In L. J. Cronbach & P. J. D. Drenth (Eds.) Mental Tests and Cultural Adaptation. The Hague: Mouton, 1972. Pp. 349-354.

Footnote

[1] I am indebted to Jane R. Mercer for the WISC-R data and the SES ratings. They have been described in detail in Jensen & Figueroa (in press).