Arthur R. Jensen

TEST BIAS AND CONSTRUCT VALIDITY

Recent research using several indices of cultural bias shows no significant black/white bias in any of a number of widely used tests of intelligence.

Most psychologists are 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 typical.

"IQ tests are Anglo-centric; 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."

"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."

"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 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."

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

The main themes in these criticisms of mental tests are:

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 between 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?

Just where do tests show differences and how big are these differences? I have been able to examine this question with a number of different intellectual tests, using very large samples of schoolchildren in California. Because of its familiarity, the Wechsler Intelligence Scale for Children–Revised (WISC-R) is a good example, with data on full scale IQs of more than 600 whites and 600 blacks representing a random sample of California schoolchildren, ages 5 to 12.[1]

Table 1 shows an analysis of variance, with estimated percentages of total variance attributable to each of the sources. The figures easiest to grasp are those in the last column, giving the average absolute difference in IQ. For these children, we based a 10-point scale of socioeconomic class on parental occupation. The average IQ differences between all possible comparisons of the 10 social classes (within each racial group) was only six IQ points. (The largest SES difference was 26 IQ points in the whites and 12 IQ points in the blacks.)

The average race difference, independent of socioeconomic status, 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.[2] If the Wechsler IQ test is so culturally biased, as some 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 between siblings 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 nine IQ points, which is 50% greater than the average difference *between* social classes.

In short, the notion that IQ tests discriminate largely in terms of race or social class is just a myth. The IQ shows as much or more difference among children in the same family, sharing the same parents and cultural and linguistic background, as between racial or social class groups.

Criteria of Cultural Bias

In discussing bias, we must first distinguish clearly 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

ARTHUR R. JENSEN is professor of educational psychology and research psychologist, Institute of Human Learning, University of California, Berkeley. This article is drawn from an address he delivered at the September, 1975, meeting of the American Psychological Association in Chicago.

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. A test item requiring the respondent to name three parks in New York City is, in this sense, more culture loaded than the question, "How many 10-cent candy bars 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. It is an empirical matter. There is no such thing as test bias in the abstract. The determination of bias involves a specific test used in two or more specific populations. 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 culture 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 this: Do the intercepts and slopes of the regression of criterion measures on test scores differ appreciably for the two populations in question? In other words, do the test scores predict equally well for both groups? When the regression lines of two groups have different intercepts and slopes, a person's predicted performance on the criterion – job, school, etc. – will be influenced by his group membership, and test scores are not adequate predictors. An unbiased test, on the other hand, is color-blind. Its prediction of a person's future scholastic or job performance, based on scores, is as accurate for blacks as for whites.

Research on this criterion of test bias is unequivocal. There is a negligible difference in the slopes and intercepts of regression lines for whites and blacks on standard tests. A single regression equation predicts equally well for both racial groups.[3] 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 outperform blacks on the criterion. In other words, the tests tend to *over*predict blacks' performance on the criterion, which gives blacks a selection advantage. 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 borne out empirically 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.

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	6
	22	
Race (within social classes)	14	12
Between families (within race and social class)	29	9
	73	
Within families (siblings)	44	12
Measurement error	5	4
Total sample	100	17

Sample size: whites = 622; blacks = 622.

Extremes of Culture Loading

First, let us contrast two tests 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 consist of 60 plates, each consisting of a pattern of geometric designs 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 3-year-olds up to a level of difficulty beyond the capacity of the average adult.

Both of these tests were individually administered to about 600 white and 400 black children, ages 6 to 12, in California schools.[4] The two groups show a 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 instruments 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 .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.)

"A variety of . . . tests have shown the same sort of thing; that is, black/white differences in test performances can be closely simulated, quantitatively and qualitatively, by comparing groups of younger and older white children."

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.

Rank Order of Item Difficulty. 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 Ps within each racial group.)

On the Peabody test, the corrected 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.

This was found not to be the case when Peabody test scores of white schoolchildren in London, England, were compared with scores of age-matched white children in California. 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. Words like bronco, thermos, and caboose, for example, are unfamiliar even to most adults in England, though they are of only moderate difficulty for American schoolchildren. Obviously the linguistic backgrounds of Londoners and Californians differ very much more than those of whites and blacks residing in California. The London children, however, also found certain words much easier (e.g., bannister), while some were more difficult, so that the overall differences average out and the English and the American 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 about the same as for whites.

If the Peabody Picture Vocabulary Test were really to reflect 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 English and Americans. But we find no difference between California 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 only at 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. The correlation of P decrements across two groups is a most sensitive index of group similarity.

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 so alike in both the black and white groups. It would seem even 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 Raven Items. Are verbal tests more biased than nonverbal? The small differences between the Peabody 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 matched Peabody and Raven items for difficulty in the white group. For each of 35 Raven items we found a Peabody item with the same percent passing. If the culture-loaded Peabody items were more biased against blacks than the culture-reduced Raven items, then we should expect blacks to obtain lower scores on the Peabody than on the Raven when the difficulties of the two sets of items are matched in the white group. It turns out that blacks show no significant difference between the Raven and Peabody scores. Raven and Peabody items matched for difficulty in the white group are thereby also matched for difficulty in the black group.

The same analysis was done for a Mexican-American group. 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.

Racial Group Item Discriminabilities

In both the Peabody and the Raven we compared *a*) the point-biserial correlations between single items and total score within each racial group, and *b*) the correlations (phi coefficients) 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 individual 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 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 nonchance fashion over the multiple-choice distractors for each item *in the same proportions* for whites and blacks. In Raven's Matrices there were several significant exceptions to this finding: 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* X *items* interaction. The size of this 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% of the total variance in both the Peabody and the Raven.

We found that we could closely *simulate*, within the margin of sampling error, this whole analysis of variance, with all of its main effects and all of 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 the features of the analysis of variance on the two racial groups. There is 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* X *items* interaction was also about 1% of the total 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* X *items* interaction became quite nonsignificant and accounted for less than .2% of the total variance.

In the light of these findings, to maintain that these tests are culturally biased with respect to black/white comparisons, one would have to argue that the *cultural differences* between California blacks and whites closely simulate *age differences* within the white group, for such a diversity of indices as rank order of item difficulties, *P* decrements, inter-item correlations, choice of distractors, and item factor loadings on the first principal component – on tests as diverse as picture vocabulary and progressive matrices! Such an argument strikes me as quite implausible.

A variety of other tests have shown the same sort of thing; that is, black/white differences in test performance can be closely simulated, quantitatively and qualitatively, by comparing groups of younger and older white children. This has been shown for developmental tests such as Piagetian conservation tests, copying simple geometric designs, and free-choice preferences for matching stimuli on the basis of color, form, size, and number preferences which change with age.[5]

Internal Bias of Other Tests

The types of analysis described above have been applied to other tests as well, all with highly similar results. But certain 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 3½ 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 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.[6]

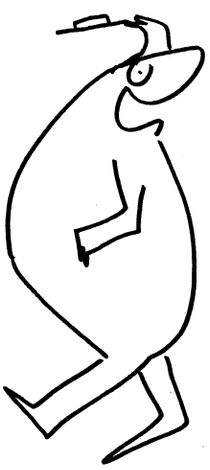
Note three important points: 1) we are dealing with only a restricted portion of the Stanford-Binet test (16 items from year III-6 through IV-6), 2) all the children are within a one-year age interval, and 3) all are preschoolers – they haven't yet been exposed to the common culture of public schooling.

The rank-order correlation between the blacks and whites in the percent passing each of these 16 Stanford-Binet items turns out to be .99 (without correction for attenuation). The Pearson correlation between the *P* values of blacks and whites is .96.

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

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 David Wechsler himself was confronted by this claim in an interview with Dan Rather on a recent CBS-TV program, "The IQ Myth."

After seeing the CBS program, a psychology graduate student, Frank Miele, looked up the item statistics on this and other WISC items. He obtained WISC tests on large samples of age-matched white and black schoolchildren 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



"(A) Good morning (B) Good evening (C) None of these."

"It is a good morning. True or false."

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 *easier* 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 appreciably 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 five black and five white psychologists could sort out the eight most and the eight 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 from a factor analysis of all the item intercorrelations within each racial group that 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 have sent me correlational data on 541 white and 237 black children in Georgia schools. All the children were given 29 cognitive tests of great variety – verbal, numerical, spatial, nonverbal reasoning, form board, vocabulary, arithmetic, spelling. The tests were taken from several different standard batteries.

"Not difficulty per se, but *complexity* is the key to *g* [general intelligence]. Items that require some active mental manipulation, some conscious mental transformation of the input, . . . are the most *g*-loaded items."

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 strong evidence that the *g* factor in this large battery of diverse tests is the same *g* for blacks as for whites.

Paul Nichols intercorrelated seven of the subtests of the Wechsler Intelligence Scale for Children 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.[7] 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 California 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 close to 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 these various cognitive tests were culturally biased for these two populations, it seems improbable that the magnitude of the bias would be so uniform over all types of tests that they would all have the same pattern of *g* loadings (within the margin of sampling error) in black and white populations. These various tests, all with substantial loadings on *g*, yield no evidence of differential cultural biases in American blacks and whites.

What Is the Nature of *g*?

What is this *g* factor that all complex 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 is 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 on the average will differ less in performance on tasks involving lesser cognitive complexity than on tasks in-

volving 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 that calls for some information processing.

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 or complex RT.[8] In these experiments, each person acts as his own control, and 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 two-choice, four-choice, or eight-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; therefore, in accord with our hypothesis, choice RT shows significant correlations with IQ and with race, while simple RT does not.

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 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.

Richard Figueroa and I tested these predictions in age-matched samples of 622 blacks and 622 whites randomly drawn from California schools.[9]

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 *within* each racial group. We also found that the difference between whites and blacks in *backward* memory span is more than twice as large as the difference in *forward* 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.

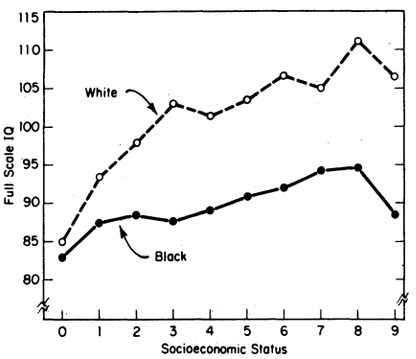


Figure 1. WISC-R Full Scale IQ of black (N = 622) and white (N = 622) samples as a function of socioeconomic status as measured by Duncan's Index of SES.

Figure 1 shows the total WISC IQs as a function of race and Duncan's index of socioeconomic status.

Figure 2 shows forward and backward digit-span scores as a function of race and SES. (The interaction of race X forward versus backward span is significant beyond the .001 level.)

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 backward digit span. I do not know of any hypothesis invoking cultural bias in the Wechsler tests that would have predicted either of these interesting psychological phenomena.

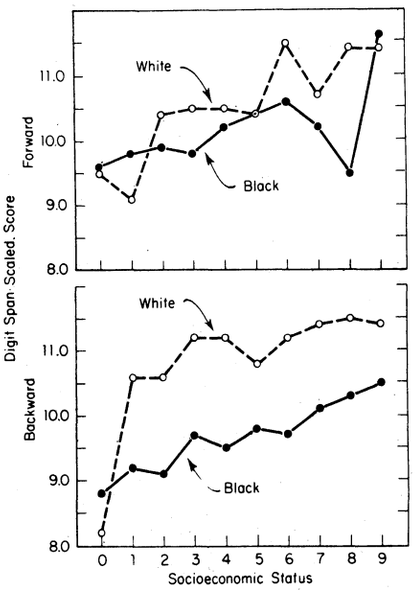


Figure 2. WISC-R Forward and Backward Digit Span scaled scores ($\chi = 10$, $\sigma = 3$) of black and white samples as a function of socioeconomic status.

Conclusion

The several 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 – show practically no evidence of differential culture bias for blacks and whites. They behave statis-

tically much the same in both racial groups and perform 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.* □

---

1. I am indebted to Jane R. Mercer for the WISC-R data and the SES ratings. They have been described in detail in A. R. Jensen and R. A. Figueroa, "Forward and Backward Digit-Span Interaction with Race and IQ," *Journal of Educational Psychology*, December, 1975, pp. 882-93.

2. The percentages of variance were estimated as follows: First, the following correlations were obtained: Race x SES = .4381; Race x IQ = .4955; SES x IQ = .4355. Then partial correlations were obtained, partialing out SES from Race x IQ, and Race from SES x IQ, yielding (Race x IQ)/SES = .3765; (SES x IQ)/Race = .2797.

The proportion of IQ variance attributable to Race and SES independently of one another is the square of the partial correlations, i.e., .14 for Race and .08 for SES.

The WISC-R gives .95 as the test-retest (one-month interval) reliability of Full Scale IQ in the age range of the present sample. This means there is 5% measurement error. Thus 14% + 8% + 5% = 27%, leaving 73% for variance Between Families (within racial and SES groups) and Within Families. By variance Between Families is meant the interfamily variability between the means of the siblings. Within Families variance is variability among siblings within the same family. The Between and Within Families variances were determined from my study of sibling correlations in large samples of whites and blacks on a highly comparable IQ scale (Lorge-Thorndike IQ), which was .43 in both racial groups. This means that 43% of the variance within racial groups is attributable to differences between families, which includes SES differences. The remainder of the variance is due to variance *within* families and error variance. This means that if we exclude variance due to race (14%) we are left with 86%, and the sum of the SES variance plus variance Between Families (*within* SES groups) divided by 86% must be .43, i.e., we solve (8% + x)/86% .43, so x = 20%, which is the percent of variance between families within SES and racial groups. The remainder of 44% is the variance within families.

3. T. A. Cleary, L. G. Humphreys, S. A. Kendrick, A. Wesman, "Educational Use of Tests with Disadvantaged Students," *American Psychologist*, January, 1975, pp. 15-41; L. G. Humphreys, "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; and R. L. Linn, "Fair Test Use in Selection," *Review of Educational Research*, Spring, 1973, pp. 139-161.

4. For full details, see A. R. Jensen, "How Biased Are Culture-Loaded Tests?" *Genetic Psychology Monographs*, November, 1974, pp. 185-244.

5. A. R. Jensen, "Race and Mental Ability," in J. F. Ebling, ed., *Racial Variation in Man* (New York: Academic Press, 1975).

6. A. R. Jensen and R. A. Figueroa, "Forward and Backward Digit-Span Interaction with Race and IQ," op. cit.; Paul L. Nichols, "The Effects of Heredity and Environment on Intelligence Test Performance in 4- and 7-Year-Old White and Negro Sibling Pairs," doctoral dissertation, University of Minnesota, 1972.

7. Nichols, op. cit.

8. J. J. Bosco, "Social Class and the Processing of Visual Information," Final Report Project No. 9-3-041, Contract No. OEG-5-9-325041-0034 (010) (Washington, D.C.: Office of Education, U.S. Department of Health, Education, and Welfare, May, 1970); A. R. Jensen, "Race and Mental Ability," op. cit.; C. E. Noble, "Race, Reality, and Experimental Psychology," *Perspectives in Biology and Medicine*, Autumn, 1969, pp. 10-30; and Y. Poortinga, "A Comparison of African and European Students in Simple Auditory and Visual Tasks," in L. J. Cronbach and P. J. Drenth, eds., *Mental Tests and Cultural Adaptation* (The Hague: Mouton, 1972), pp. 349-54.

9. Jensen and Figueroa, "Forward and Backward Digit-Span Interaction with Race and IQ," op. cit. □

Graduate Students Find Studies Disappointing and Damaging

≫ Many graduate students in U.S. institutions of higher education find their studies intellectually disappointing and emotionally damaging, according to a two-year survey by researchers at Berkeley's Wright Institute.

Many students "find their lives crammed, their moods serious if not grim, and their energies beset by relentless requirements and even busywork, all of which make graduate school at times more resemble military drill than the exercise of man's most intellectual and imaginative capacities," the researchers say in the *Chronicle of Higher Education*.

The researchers, Joseph Katz of the State University of New York at Stony Brook and Rodney Hartnett of the Educational Testing Service, based their conclusions on in-depth interviews with more than 100 graduate students in the Berkeley area, questionnaires returned by more than 700 graduate students at four universities, and on nationwide data gathered by the ETS.

Financed by grants from the Lilly Endowment and the National Institute of Education, the report criticizes graduate departments of education for failing to train students for their role as teachers of undergraduates. Instead, the researchers claim that graduate students are "taught to neglect teaching if not to have contempt for it." In addition, graduate schools have not adjusted to economic hard times and declining job markets for their students, the *Chronicle* reports.

"Apparently the leadership in graduate education is taking very little initiative in pressing for a rethinking of the goals and purposes of most graduate programs, in spite of clear evidence that the old assumptions and the old attitudes are no longer adequate," the report claims.

The two researchers call for changes in the structure of graduate education to make the experience less traumatic for students. Students arrive at graduate school with expectations that are quickly thwarted: 1) They hope to join a community of scholars, but instead are pushed into "relative intellectual isolation and concentrating in a narrow specialty." 2) They expect lively interactions, but find "competitive atmospheres and inadequate opportunities for working with others." 3) They find "access to professors limited" and at times are subjected to "demeaning" treatment. They are treated like college freshmen, not like members of a community of intellectual peers.

The authors maintain that many of these problems, among them "a loss of theoretical breadth, community of inquiry, and civility," have been created simply by the growth in the number of graduate students and in the size of graduate departments.

Katz and Hartnett offer a number of recommendations to remedy graduate students' problems. They urge "greater equalization of the flow of information between prospective graduate students and graduate departments." They also recommend that faculty members become more sensitive to the emotional problems of graduate students, as well as that the supply of graduate and professional students be limited to "avoid the creation of a high-class intellectual proletariat," that limits be placed on the number of years students spend in graduate education, and that teaching be made a "prestigious part of graduate training."