DOCUMENT RESUME

ED 062 483 UD 012 361

AUTHOR Jensen, Arthur R.
TITLE Improving the Assessment of Intelligence.
PUB DATE 28 Apr 72
NOTE 9p.; paper presented in symposium, "The Theory of Intelligence," at the Western Psychological Association Annual Meeting, Portland, Ore., April 28, 1972

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS Academic Achievement; Cultural Factors; *Educational Diagnosis; Factor Analysis; Intelligence Differences; *Intelligence Tests; Item Analysis; Matched Groups; Racial Differences; Response Style (Tests); *Test Bias; *Testing Problems; Test Interpretation; *Test Validity
IDENTIFIERS California

ABSTRACT

Contrary to popular opinion, it is very difficult to find any objective evidence of culture bias that could account for social class and racial differences in performance on current standard tests of intelligence, even those like the Peabody Picture Vocabulary Test (PPVT), which give the appearance of being highly culture-loaded. They may be culture-loaded, but there is no evidence one has been able to find that the culture-loading differentially affects the performance of Negro and white children. Difference in mean score cannot be a criterion of culture bias. One must seek other evidence. The following types of evidence have been examined in the PPVT, the Lorge-Thorndike Intelligence Test, and Raven's Progressive Matrices, these studies having involved very large samples of Negro and white children in several California school districts: (1) The rank order of item difficulty is virtually the same for Negroes and whites; (2) The matrix of item intercorrelations and the factor structure of these tests is not significantly different for white and Negro samples when these are roughly matched for mental age or total score; (3) In multiple-choice tests, there is no systematic or significant racial difference in the choice of distractors on those items that are answered "wrong"; and, (4) The intelligence tests show essentially the same correlations with scholastic achievement in Negro and white samples. (Author/JM)

Improving the Assessment of Intelligence*

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

SUMMARY

At a time when intelligence and aptitude testing are under fire for largely ideological and political reasons, it behooves psychologists to re-examine their theories of mental abilities and the psychometric techniques for assessing them. The present paper briefly summarizes three closely interrelated lines of investigation that I have been pursuing and that are germane to the current controversies about intelligence testing.

Intelligence Only a Part of Mental Ability

What we technically call intelligence, or the g factor common to nearly all complex tests of ability, is only a part of the total spectrum of human abilities. It has been the most strongly emphasized ability in our culture because it is so closely related to scholastic performance and to occupational and consequently general socioeconomic status.

The total domain of mental abilities may be sliced up by factor analysis or other multivariate techniques in a variety of ways. There is Spearman's two-factor model, Burt's hierarchical model, Thurstone's primary mental abilities, Guilford's structure of intellect model, Cattell's fluid and crystalized intelligence, to name a few of the ways in which the

*Paper presented in symposium, "The Theory of Intelligence," at the annual convention of the Western Psychological Association, Portland, Oregon. April 28, 1972.

abilities domain has been represented. A division of abilities that I have been concerned with involves a distinction between what I now call Level I and Level II abilities. Level I abilities involve simple learning and association, the registering, retention, and retrieval of inputs. It involves very little or no mental manipulation of the input. Level I can be thought of mainly as rote learning and memory. Level II, on the other hand, inplies mental manipulation, the ability to deal with complexity, information processing, and the active relating and comparing of present inputs with stored past inputs. It involves the imposing of cognitive structures upon sensory inputs. Level II is the g factor of intelligence, particularly fluid intelligence. Level I is best measured by memory span for digits, serial rote learning, paired associate learning, free recall memory, and trial-and-error selective learning. Level II is best measured by tests of fluid intelligence such as Raven's Progressive Matrices and Cattell's Culture Fair Tests of g.

The interesting thing about the Level I - Level II distinction from my standpoint is that it is the only broad division of the mental abilities domain that clearly interacts with social class and particularly race. While social class and racial (i.e., white-Negro) differences are prominent on all Level II tests, they are small or negligible on Level I tests. Moreover, there are racial differences in the extent of correlation between Level I and Level II abilities, the correlation being higher in white and lower in Negro populations. Since an individual's overall social competence may be related to both Level I and Level II, I believe this finding has potentially important and useful implications for the assessment of abilities in those population groups which generally score low, on the average, on Level II abilities. It means, among other things, for example, that low IQ

Negro children may be less handicapped overall than low IQ white children in Level I functions. The extent to which Level I ability contributes to overall competence in our society is still obscure. We do know that Level II ability is of prime importance in scholastic performance as schools are presently constituted, and for success in certain occupations in which performance is dependent upon scholastic skills. At least we have discovered one part of the ability domain which shows no appreciable racial difference. It remains to determine the significance of these Level I abilities in practical affairs and their possible utilization in scholastic learning and in the development of useful occupational skills.

Culture Bias in Standard Tests

Contrary to the popular mythology in this field, it is very difficult to find any objective evidence of culture bias that could account for social class and racial differences in performance on current standard tests of intelligence, even those, like the Peabody Picture Vocabulary Test (PPVT), which give the appearance of being highly culture-loaded. They may be culture-loaded, but there is no evidence we have been able to find that the culture-loading differentially affects the performance of Negro and white children. Difference in mean score cannot be a criterion of culture bias. One must seek other evidence. We have examined the following types of evidence of culture-bias in the PPVT, the Lorge-Thorndike Intelligence Test (which is most widely used in California schools), and Raven's Progressive Matrices. These studies have involved very large samples of Negro and white children in several California school districts.

Rank order of item difficulty (p values). The rank order of the percent passing each item is virtually the same for Negroes and whites. The correlations between p values for these tests are all above .95, averaging

.97. In this respect, the two racial groups are more alike than are boys and girls within each race. In other words, the cultural biases in the test are more apparent with respect to sex differences than with respect to race differences. (The sexes do not differ appreciably in mean score, however, while the racial groups differ about one standard deviation, or 15 IQ points, on the average.)

Item intercorrelations and factor structure. The matrix of item intercorrelations and the factor structure of these tests is not significantly different for white and Negro samples when these are roughly matched for mental age or total score. These properties of the data, for example, do not in the least distinguish between 4th grade white children and 6th grade Negro children. Yet they distinguish between 5th grade and 6th grade Negro children and 5th grade and 6th grade white children. A culture-bias hypothesis would predict greater Negro-white differences than adjacent grade differences in item intercorrelations. The findings, on the other hand, are more consistent with a developmental lag hypothesis.

Choice of distractors. In multiple-choice tests (as the Lorge-Thorndike, PPVT and Raven are), there is no systematic or significant racial difference in the choice of distractors on those items that are answered "wrong." A special scoring key was made up so as to score as correct whatever response is given by the largest number of children in the Negro sample. When the tests are scored by this key, the Negro sample still averages lower than the white sample.

Scales based on subgroups of items which discriminate either least between Negroes and whites or discriminate most are correlated with each other over .90 (approximately the reliability of the test), showing that the two types of items are measuring the same ability.

Correlations with Scholastic Achievement. The intelligence tests show essentially the same correlations with scholastic achievement in Negro and white samples. When scholastic achievement is "predicted" by a multiple regression equation comprised of several intelligence tests, adding race (white vs. Negro) to the multiple prediction equation does not increase the multiple correlation with scholastic achievement. Thus, the predictive validity of the IQ test is the same for Negroes and whites. Negroes and whites with the same IQ perform equally well in school.

In short, none of our analyses reveals any racial differences other than the number of items gotten right. There seems to be no good reason to believe that these tests behave any differently for Negroes than for whites.

Heritability and Sibling Correlations. The sibling correlations on 16 ability tests were examined in large Negro and white samples. They are very similar, as indicated by a correlation of .71 between the sibling correlations on each test for Negroes and whites. The average difference between siblings on each test does not differ significantly for Negroes and whites.

When estimates of the heritability (i.e., the proportion of genetic variance in test scores) of the various tests are correlated with the magnitude of the mean white-Negro difference on the tests, the correlation is positive (.80 for whites, .61 for Negroes). In other words, those tests which are least sensitive to environmental influences (i.e., high heritability) in general show the largest white-Negro differences and those tests which are most sensitive to environmental influences (i.e., low heritability) show the smallest Negro-white differences. This outcome is just the opposite of what one would expect from a culture-bias or environmental hypothesis of the cause of the racial difference. This study has been repeated by

other investigators using a different set of tests, and the results are essentially the same, i.e., a strong positive correlation between tests' heritability and the magnitude of the white-Negro difference.

Those who claim culture bias in current widely used tests, it seems to me, are obligated to produce some objective evidence that such bias in fact exists. I have found no evidence that it does, at least in the well-known tests we have studied.

Reaction Time as a Measure of Information Processing Capacity

Is it possible to devise a test of intelligence which not only is not culture biased but is not appreciably culture loaded (at least within Western industrialized societies)? And which has such added advantages as repeatability on the same test, applicability over a wide age range such that the test itself need not change for widely different age groups, and, perhaps most important, a test on which the subject's performance is not self-reactive, i.e., his performance on a given item does not affect his attitude toward the test or his performance on subsequent items, as might be caused by a series of difficult items which the subject fails, leaving him too discouraged or too poorly motivated to tackle further items.

Reaction time as a measure of information processing capacity may provide a test with these characteristics. Information processing capacity is a component, at least, of g or fluid intelligence; perhaps it is the essence of g. This remains to be determined. I have made a beginning by measuring information processing capacity by means of reaction time (RT) to stimulus situations that convey different amounts of "information" in the sense in which this term is used in information theory. Subjects are required to turn off a green light as fast as possible after it goes "on"

by pressing a button directly adjacent to the light. The amount of information is varied by presenting different numbers of light/button alternatives. The subject never has to turn off more than one light. The number of bits of information conveyed equals the logarithm, to the base 2, of the number of light/button alternatives.

The apparatus for measuring the subject's RT and MT consists of a panel, 13" x 17", painted flat black, and tilted at a 30° angle. At the lower center of the panel is a red pushbutton, ½" in diameter, called the "home" button. Arranged in a semi-circle above the "home" button are eight red pushbuttons, all equidistant (6") from the "home" button. Half an inch above each button (except the "home" button) is a ½" faceted green light. Different flat black panels can be fastened over the whole array so as to expose arrays having either 1, 2, 4, 6, or 8 light/button combinations.

The subject is instructed to place the index finger on the "home" button; then an auditory warning signal is sounded (a high-pitched tone of 1 sec. duration), followed, after a continuous random interval of from 1 to 4 seconds, by one of the green lights going "on," which the subject must turn off as quickly as possible by touching the sensitive microswitch button directly under it. RT is the time the subject takes to remove his finger from the "home" button after the green light goes on. MT is the interval between removing the finger from the "home" button and touching the button which turns off the green light. RT and MT on each trial were registered in milliseconds (ms) by two electronic timers.

Every subject was given a total of 30 trials on each of the five arrays (i.e., 1, 2, 4, 6, or 8 light/button alternatives, corresponding to 0, 1, 2, 2.58, and 3 bits of information, respectively). The particular light that went on in each trial was random and hence unpredictable by the subject.

Reliabilities for RT and MT are all above .90 for each of the five conditions. The task is exceedingly obvious and easy for all subjects (we haven't tested any younger than grade 3) and there is little or no subjective sense of increase in task difficulty as subjects move from the 1-button to the 8-button test.

For virtually all subjects the RT is a linear increasing function of the bits of information. MT shows no increase as a function of bits. The correlation between RT and MT is only about .30.

RT correlates significantly and substantially with a good test of g, Raven's Progressive Matrices. A multiple R with Matrices scores based on RT to each of the five tests (1, 2, 4, 6, and 8 light/button alternatives) was over .60. The addition of MT to the regression equation does not boost the multiple R appreciably. The slope of the RT function with increasing information load is also significantly correlated with intelligence test scores. Brighter subjects show a smaller increase in RT as the information processing demands of the task increase. The slope of RT over bits is likely to prove the most important aspect of the test, since it is this feature, rather than simple RT per se, which most clearly reflects information processing capacity.

This research is still at the stage of investigating the psychometric correlates and the factor composition of the RT and MT measures. The technique has not been applied to the study of population differences and will not be until much more is definitely known about its psychometric properties.

ADDENDUM

The results reported in the attached Summary are based on a pilot study of only 39 Ss. A subsequent study has just been completed based on 160 middle-class and upper-middle-class white children in grades 4 through 6. This group is considerably more homogeneous in abilities than the subjects used in the pilot study (i.e., Jensen & Munro).

The results of this larger study will later be published in detail in an appropriate psychological journal. The results can be summarized briefly as follows. The multiple correlations between the reaction time (RT) plus movement time (MT) measures (in milliseconds) and four psychometric test scores are as follows:

Lorge-Thorndike Intelligence - Verbal $R = -.356$ (-.322)
Lorge-Thorndike Intelligence - Nonverbal $R = -.333$ (-.303)
Raven's Standard Progressive Matrices $R = -.330$ (-.267)
Memory for Digits $R = -.270$ (-.214)

The figures in parentheses are the correlations when chronological age (in months) is partialled out. The P values of the Rs with age partialled out are all less than the .01 level except for the Memory test, for which $P < .02$.

A surprising finding is that the slope of the regression of RT on BITs of information does not show a statistically significant correlation with any of the psychometric test scores, either before or after CA is partialled out. This is most puzzling because it was expected that the slope rather than RT per se should be negatively correlated with the intelligence measures, since the linear increase in RT as a function of BITs of information that has to be processed in making the RT response would seem to be related to some aspect of mental ability. The slope measure for individual Ss has a reliability coefficient of 0.84. The slope is close to linear for all Ss, and for the group as a whole it is perfectly linear. Yet this slope measure for individuals shows totally nonsignificant correlations with the psychometric test scores.

It appears that RT and MT have a small but significant component of variance in common with standard psychometric measures of intelligence. Why the slope of RT as a function of amount of information does not correlate significantly with intelligence is a mystery that remains to be solved.

It should be emphasized that RT and MT account for only a small fraction of the ability test variance. RT and MT therefore cannot be used for any practical diagnostic or selection purposes involving decisions about individuals.

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