DOCUMENT RESUME

ED 080 644 UD 013 764

AUTHOR Jensen, Arthur R.
TITLE How Biased Are Culture-Loaded Tests?
PUB DATE 73
NOTE 85p.

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Caucasian Students; Cultural Differences; Cultural Factors; *Culture Free Tests; *Ethnic Groups; *Mexican Americans; *Negro Students; Racial Differences; Testing
IDENTIFIERS California; Peabody Picture Vocabulary Test; Ravens Progressive Matrices

ABSTRACT

The culture loaded Peabody Picture Vocabulary Test (PPVT) and the culture reduced Raven's Progressive Matrices (Colored and Standard forms) were examined and compared for large samples of white, black, and Chicano school children, K-8, in three California school districts. On both the PPVT and the Raven's the three ethnic groups show large mean differences but very little difference in the rank order of item difficulties, relative difficulty of adjacent items, the loadings of items on the first principal component, and the choice of distractors for incorrect responses. On both tests, groups of culturally homogeneous younger and older white children (separated by two years) perfectly simulated the white/Negro differences in Ethnic Group x Item interactions and choice of error distractors in the Raven's. Certain expectations from a culture bias hypothesis were borne out only for PPVT in the Mexican group. Unless the unlikely and empirically unsubstantiated assumption is made that culture bias affects all kinds of test items about equally, the various item analyses of the present studies lend no support to the proposition that either the PPVT or the Raven's is a culturally biased test for blacks. (Author/RJ)

How Biased Are Culture-Loaded Tests?

Arthur R. Jensen

University of California, Berkeley

ABSTRACT

The culture-loaded Peabody Picture Vocabulary Test (PPVT) and the culture-reduced Raven's Progressive Matrices(Colored and Standard forms) were examined and compared in terms of various internal criteria of culture bias in large representative samples of white, Negro, and Mexican-American school children, from kindergarten through 8th grade, in three California school districts. On both the PPVT and the Raven the three ethnic groups, which show large mean differences, show very little difference in the rank order of item difficulties, the relative difficulty of adjacent items, the loadings of items on the first principal component, and the choice of distractors for incorrect responses. Analysis of variance revealed very small Ethnic Group × Items interaction, but a sensitive index of item bias derived from ANOVA indicates that the Raven is considerably less biased than the PPVT, especially in the Mexican group. The Groups × Items interaction was shown to be attributable largely to differences in mental maturity. On both tests groups of culturally homogeneous younger and old c white children (separated by 2 years) perfectly simulated the White/Negro differences in Group × Item interactions and choice of error distractors in the Raven. Certain expectations from a culture bias hypothesis were borne out only for the PPVT in the Mexican group. Unless the unlikely and empirically unsubstantiated assumption is made that culture bias affects all kinds of test items about equally, the various item analyses of the present studies lend no support to the proposition that either the PPVT or the Raven is a culturally biased test for Negroes.

How Biased Are Culture-Loaded Tests?

Arthur R. Jensen

University of California, Berkeley

Standard tests of intelligence and scholastic aptitude, it is often claimed, are culturally biased so as to favor white subjects of middle and upper-middle-class backgrounds and to disfavor subjects of lower socioeconomic status, especially certain ethnic and racial minorities. Such culture bias is often regarded as the main explanation for mean test score differences between particular subpopulations within the United States.

In researching the validity of these claims, investigators have had to establish various objective criteria of culture bias in tests, so that its existence and magnitude might be assessed. It seems to be agreed upon by nearly all psychometric researchers that the presence of population differences in the distribution of test scores is by itself not a proper criterion for judging test bias. The argument that any test which shows group mean differences is therefore biased obviously begs the question.

The psychometrically defensible criteria of test bias that have been proposed in the literature fall into two classes: external and internal. The first is certainly the more important from the standpoint of practical prediction. The second, however, may be even more directly relevant to many current popular criticisms of mental tests on the grounds that they are culturally loaded, therefore culturally biased. The fact that they may meet certain criteria of external validity may be attributed to culture

bias in the criterion. Whether such bias is "fair" or "unfair" to the members of one or another group is another matter which must be argued on still other grounds, usually involving matters of social policy rather than psychometrics.

The external criteria of test bias have been the most thoroughly discussed and studied (e.g., Cleary, 1968, Darlington, 1971; Humphreys, 1973; Jensen, 1968, Linn, 1973; Thorndike, 1971). External evidence for bias is based essentially on the regression of a criterion measure on test scores in the two (or more) groups under consideration. If the intercepts and slopes of the regressions in the two groups do not differ significantly (or by more than some predetermined magnitude), the test is regarded as "fair" or unbiased with respect to its predictive validity for the criterion in question. The above cited references all explicate this approach and its variations and interpretations. The bulk of related empirical findings involve comparisons of white and Negro samples. Concerning these studies, Humphreys (1973, p. 59) stated: "When the literature reporting regression comparisons is summarized, the following conclusion seems warranted: there is relatively little difference in the slopes or intercepts of regression lines as a function of the demographic groups that have been studied. Use of a single regression equation for these groups leads to no substantial degree of unfairness in drawing inferences concerning the criteria measured." The criteria have generally been scholastic and job performance.

Internal criteria of test bias involve item analyses and particularly evidence of Groups × Items interaction. One kind of evidence of such interaction is seen when the rank order of difficulty of items (as indicated by p, the percent passing each item) is significantly different in two populations. Another evidence of interactions is seen even when the rank order

of p values is the same in both groups but the differences between the p values of adjacent items are significantly different in the two populations. Analysis of variance (ANOVA) provides an overall test of Groups × Items interaction, but confounds the two types of interaction just described, i.e., (a) based on the rank order of p values and (b) on the differences between p values of adjacent items. These a and b types of interaction are also referred to respectively as ordinal and disordinal.

The ANOVA approach to internal evidence of bias is illustrated in a study by Cleary and Hilton (1968), who examined the interactions of individual items on two forms of the Preliminary Scholastic Aptitude Test in white and Negro groups. The Race × Items interaction was statistically significant but contributed so minimally to the total variance that the authors concluded: " . . . given the stated definition of bias, the PSAT for practical purposes is not biased for the groups studied." Stanley (1969) showed that a considerable amount of this interaction was due to just a few items that were too difficult for both races and thus did not discriminate much between them. The Negroes scored rather uniformly lower than whites on most of the items.

Both external and internal criteria are important in the study of test bias. Internal criteria may in fact be a more powerful indicator of culture bias per se, while external criteria reflect any of a number of factors that can lower a test's predictive validity in a particular population. Internal criteria seem especially appropriate for investigating the hypothesis that a given test is biased for one population when the item selection and standardization were based on a different population. If the test items are culture-loaded, i.e., they call for specific information acquired in a given culture, and if the cultures of the standardization

and target groups differ with respect to the cultural information sampled by the items, this should be reflected in various internal indices of bias, such as Culture-group × Item interactions.

The claim of cultural difference is the most common criticism of standard ability tests. Thus, the Council of the Society for the Psychological Study of Social Issues (1969, p. 1039) states: "We must also recognize the limitations of present day intelligence tests. Largely developed and standardized on white, middle class children, these tests tend to be biased against black children to an unknown degree." The cultural difference model holds that intelligence test differences between blacks and whites ". . . are manifestations of a viable and well-delineated culture of the Black American. . . . Blacks and whites come from different cultural backgrounds which emphasize different learning experiences necessary for survival" (Williams, 1971, p. 65). Williams goes further: "A review of the research on comparing intellectual differences between Blacks and whites shows the results to be based almost exclusively on differences in test scores, or I.Q. Since the tests are biased in favor of middle-class whites, all previous research comparing the intellectual abilities of Blacks and whites should be rejected completely" (p. 63). It is not said by which criteria such cultural bias has been established or how its magnitude relative to other sources of test variance has been estimated. These are proper questions for study.

Mercer (1973) has helped by posing the question of culture bias somewhat more pointedly and naming specific tests which she believes most exemplify culture bias. Her position can be summarized by some direct quotes: "American I.Q. tests have, inevitably, included items and procedures which reflect the abilities and skills valued by the American core

culture. This 'core culture' consists mainly of the cultural patterns of that segment of the population consisting of white, Anglo- Saxon Protestants whose social status today has become middle and upper-middle class" (p. 66). She suggests that the low average IQ test score of minority children results primarily from lack of exposure to the Anglo core culture (p. 108).

As an example of a white-Anglo culture-biased test--the most extreme among eleven tests that were examined--Mercer points to the Peabody Picture Vocabulary Test (PPVT). That the test items are culture loaded is obvious from mere inspection. Whether they are biased, and to what extent, with respect to any given population, however, is a separate question and is the main point at issue. Merely to point out that the test is culture loaded does not of itself consittute evidence that the test is biased with respect to the populations in question. Mercer rightly notes, however, that in the PPVT "The child must be familiar with a wide variety of objects, for example, ambulance, tweezers, wasp, captain, hive, reel, idol, casserole, scholar, and observatory. He must also be able to decode the pictures to determine which one best represents such words as filing, harvesting, soldering, assistance, dissatisfaction, astonishment, and horror. In some cases, the words in the vocabulary list are not the words most commonly used in spoken English for the objects which are pictured, for example, shears, chef, cobbler, and hydrant. In the case of some adjectives, the picture is of an object which the adjective frequently modifies. For example, the correct response to the word thoroughbred is the picture of a horse" (p. 71).

Culture-Loaded and Culture-Reduced Tests

Because the PPVT is so generally conceded to be perhaps the most

obviously culture-loaded test among the more widely used measures of I.Q., it was selected for examination in the present study. No case is being made here for its validity or usefulness as a measure of intelligence. It is used in the present study only because it is so obviously "cultural" in the same sense that the quotes from the SPSSI Council, Williams, and Mercer intend this term to mean. In the present writer's opinion, the PPVT is probably much too narrow in the variety of abilities it taps (viz., recognition or receptive vocabulary) to be a good measure of general intelligence in the sense of g, i.e., the factor common to a wide variety of mental tasks. The obviously culture-loaded PPVT, however, should be an ideal instrument for the investigation of internal evidence of cultural differences and of culture bias in testing non-Anglo minorities. In the present study these are Negroes and Mexican-Americans.

Peabody Picture Vocabulary Test.--Detailed descriptions of the PPVT and its standardization are provided by Dunn (1965) and Buros (1965, pp. 820-823). Briefly, the PPVT consists of 150 plates, each with four panels containing clear-cut line drawings. (These 150 × 4 = 600 pictures were originally selected, in terms of various item-analysis criteria, from a pool of 3,885 illustrable words taken from Webster's New Collegiate Dictionary, Second Edition [1956]. 80% of the stimulus words are nouns; the rest are the present participle form of various verbs, and there are a few adjectives and adverbs.) The examiner "names" one of the four pictures on each card and the subject simply points to the appropriate picture. (The two equivalent forms of the test, A and B, use the same set of pictures but different stimulus words.) The untimed test is individually administered. No one subject is given all 150 plates. The items are

arranged in their order of difficulty in the normative sample. In giving the test, a "basal" point is established for each individual, consisting of 8 consecutive correct responses prior to the first error; all items preceding this point are assumed correct. Testing is discontinued when the subject reaches his "ceiling," which is 6 failures out of 8 consecutive responses, i.e., the expected error rate under sheer guessing. The PPVT was standardized in the late 1950s on some 4,000 white children and youths, ages 3 to 18, in and around Nashville, Tennessee.

PPVT and Thorndike-Lorge Word Frequencies.--One indication of the cultural nature of a test's item content is the degree of relationship between item difficulties (as indexed by percent passing in the normative sample) and the probability or frequency of encountering the informational content of the items in the so-called core culture. Thus the difficulty of vocabulary items may be related to frequency of exposure or usage of the words in the general population. The rank order of difficulty of the PPVT stimulus words in the normative sample were correlated with the rank order of their frequencies of occurrence (per million words) in American newspapers, magazines, and books as listed in the Thorndike-Lorge (1944) general word count. Figure 1 shows the mean frequencies within se s of 15 PPVT items. It is clear that PPVT item difficulty is very closely

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

Insert Figure 1 about here

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

related to the rarity of the words in general usage in American English.

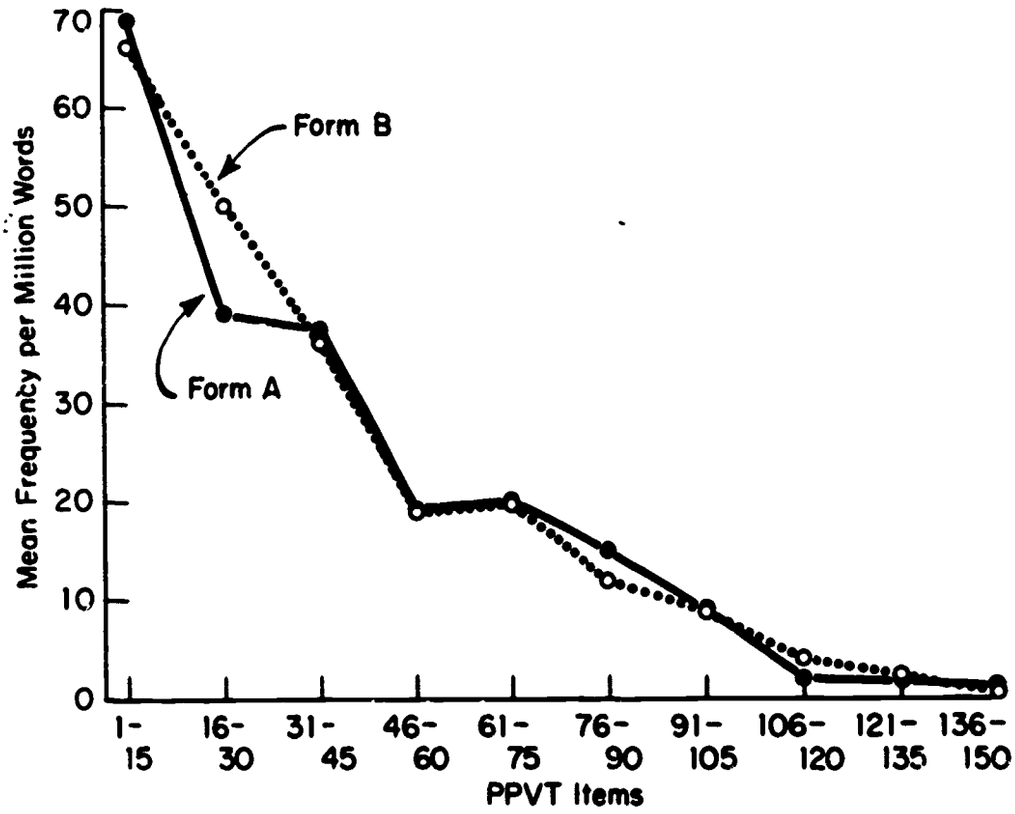

Fig. 1. Mean Thorndike-Lorge word frequency of PPVT items (in Forms A and B) as a function of item difficulty when items are ranked from 1 to 150 in p values (percent passing) based on the normative sample.

It is rarity more than the complexity of the mental processes involved that determines difficulty in the PPVT. There appears to be nothing any more difficult conceptually about culver (item 150) than about table (item 1).

It is this rarity feature of culture-loaded tests that so-called "culture-free" or "culture-fair" tests attempt to minimize. As MacArthur and Elley (1963) have suggested, such tests are better called "culture-reduced." Probably the best known and most widely used of such tests is Raven's Progressive Matrices (Raven, 1960; Buros, 1965, pp. 762-765). Such nonverbal tests are expressly designed to reduce item dependence on acquired knowledge and to keep cultural and scholastic content to a minimum while getting at basic processes of intellectual ability. Item difficulty in such tests is closely related to the complexity of the items (usually abstract figural material) and the number of elements involved in the reasoning required for the correct solution.

Thus, as the most extremely contrasting test to the PPVT on the continuum from "culture-loaded" to "culture-reduced," Raven's Progressive Matrices tests were selected for comparison with the PPVT in the present study. Two forms of the Raven were used: the Colored Progressive Matrices, for younger children, consists of 36 colored multiple-choice matrix items; the Standard Progressive Matrices, for older children and adults, consists of 60 matrix items. The items were standardized on children and adults in England. The matrix problems vary in difficulty, from the easiest, which are passed by most 3-year-olds, to the hardest, which are beyond the average adult. In both forms of the test, the items are arranged in order of difficulty within groups of 12 items, going from easy to difficult within each group, so that subjects will be less apt to become discouraged by a long

succession of difficult items as might occur if all the items were presented in order of difficulty through the entire test. It is an untimed power test; it can be individually or group-administered, and subjects are encouraged to attempt all items.

MacArthur and Elley (1963), in a study comparing verbal and culture-loaded tests with Raven's Matrices and other culture-reduced tests in a Canadian white population, found that the culture-reduced tests (a) sample the general intellectual ability factor as well or better than conventional tests, (b) show negligible loadings on verbal and numerical factors, (c) show significantly less relationship with socioeconomic status than do conventional tests, and (d) show less variation in item discrimination between social classes.

Study I. A Comparison of PPVT and Raven's Matrices in White, Negro, and Mexican-American Samples.

Tests and Subjects

Representative samples totaling 1,663 children in about equal numbers from kindergarten through sixth grade were individually administered the PPVT (Form B) and Raven's Colored Progressive Matrices in two one-hour sessions by school psychometrists (all were white) in the public schools of Riverside, California.[2] The sample sizes, by ethnic group and sex, are as follows:

White		Negro		Mexican	
Male	Female	Male	Female	Male	Female
333	305	183	198	334	310

Results

Descriptive Statistics.--The PPVT and Raven raw score means and SDs in each age group are shown in Figures 2 and 3. The overall ethnic group differences expressed in σ units, where σ is the average within-group standard deviation, are given in Table 1. The interesting feature of these

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

Insert Figures 2 and 3 about here

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

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

Insert Table 1 about here

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

comparisons is that the two minority groups are reversed in relative standing on the two tests. Though all the Mexican children in this sample spoke English predominantly, some were from bilingual homes. However, the idea that this reversal of the minority groups on PPVT and Raven is attributable simply to bilingualism or unfamiliarity with spoken English in the Mexican group should lead to the expectation of a significantly lower correlation between PPVT and Raven scores in the Mexican than in the two other groups. The fact that when age in months is controlled (i.e., partialed out) the correlation between PPVT and Raven is quite low indicates that although the two tests are measuring something in common (most probably g), they are also measuring different abilities to a more considerable extent. The relevant correlations are shown in Table 2.

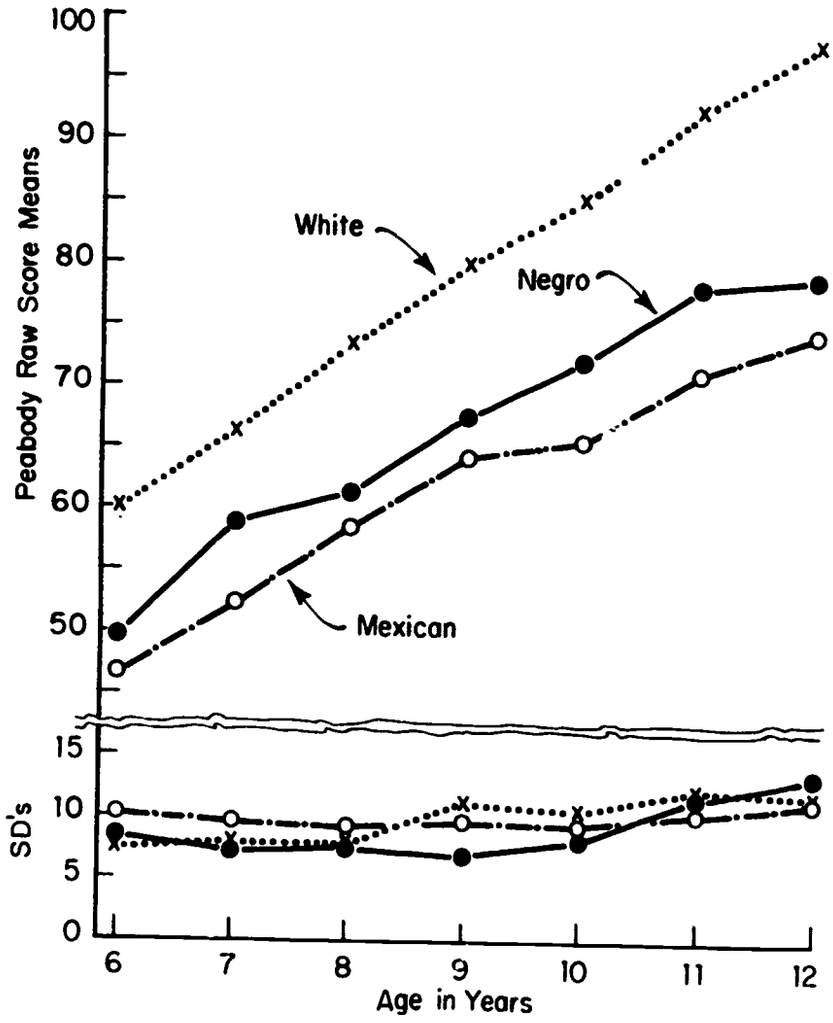

Fig. 2. PPVT raw scores as a function of age. Standard deviations (SDs) at each age are shown in lower part of graph. The ages 6, 7, etc. represent the midpoints of the intervals 5 yrs. 6 mo. - 6 yrs. 5 mo, 6 yrs. 6 mo. - 7 yrs. 5 mo., etc.

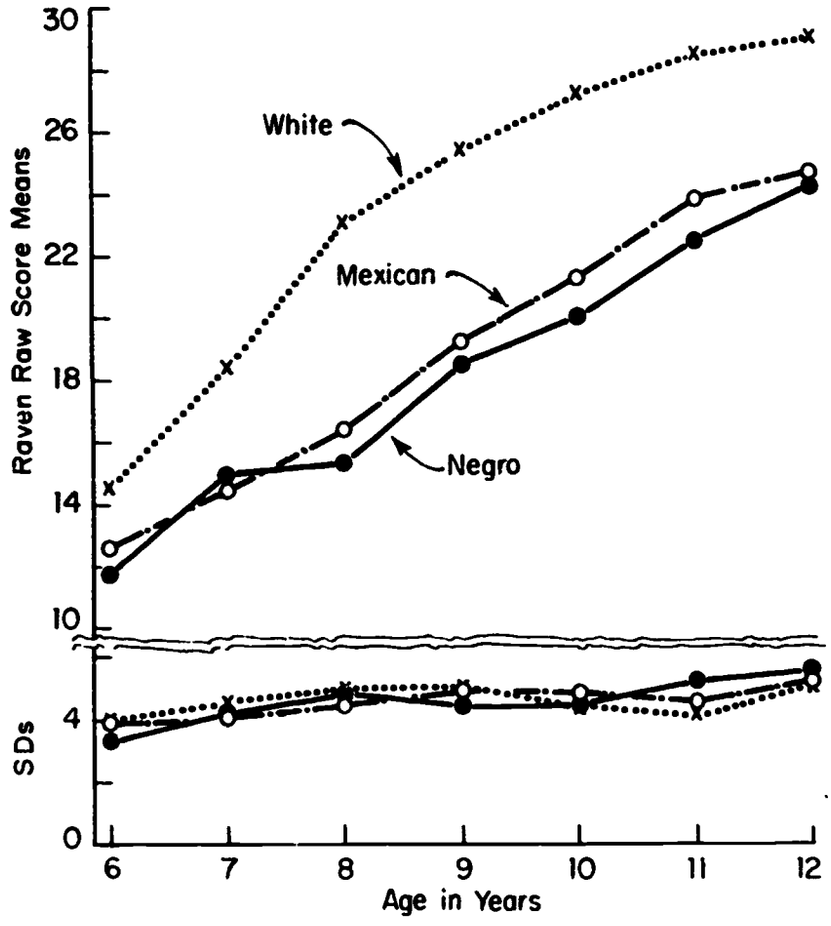

Fig. 3. Raven's Colored Progressive Matrices raw scores as a function of age.

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

Insert Table 2 about here

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

If the Raven is less culture-biased than the PPVT, one would expect that when minority and majority subjects are matched on the more culture-loaded PPVT score, the minority subjects will score higher on the presumably less culture-loaded Raven, and that when the groups are matched on the Raven, the minorities should score lower on the PPVT. These expectations can be checked in terms of the regression of each test on the other in each of the three groups. Before obtaining the regression lines, raw scores on both tests were transformed to $\underline{Z}$ scores for the entire sample, so that the group differences in the graphical presentation could be easily viewed in terms of $\underline{Z}$ scores or σ units, as shown in Figure 4. None of the regression lines departs significantly from linearity throughout the entire range of scores, and are drawn so as to include the full range of scores within each ethnic group. This can be seen to span approximately six σ in each group. The vertical arrows indicate the locations of the bivariate means for each group. An overall statistical test of coincidence of the regression lines

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

Insert Figure 4 about here

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

of the three groups in both graphs shows that they differ significantly beyond the .01 level. They differ significantly in intercepts but not in slope. The lower graph in Figure 4 entirely accords with the above-described expectation; that is, for any given Raven score, both minority

Table 2

Correlation Between Age (in months), PPVT and Raven, and Between the Tests After Age Is Partialed Out

Correlation	White	Negro	Mexican	Total
PPVT × Age	.787	.728	.671	.632
Raven × Age	.722	.660	.702	.654
PPVT × Raven	.719	.692	.667	.724
Partial r				
PPVT × Raven	.354	.412	.371	.531

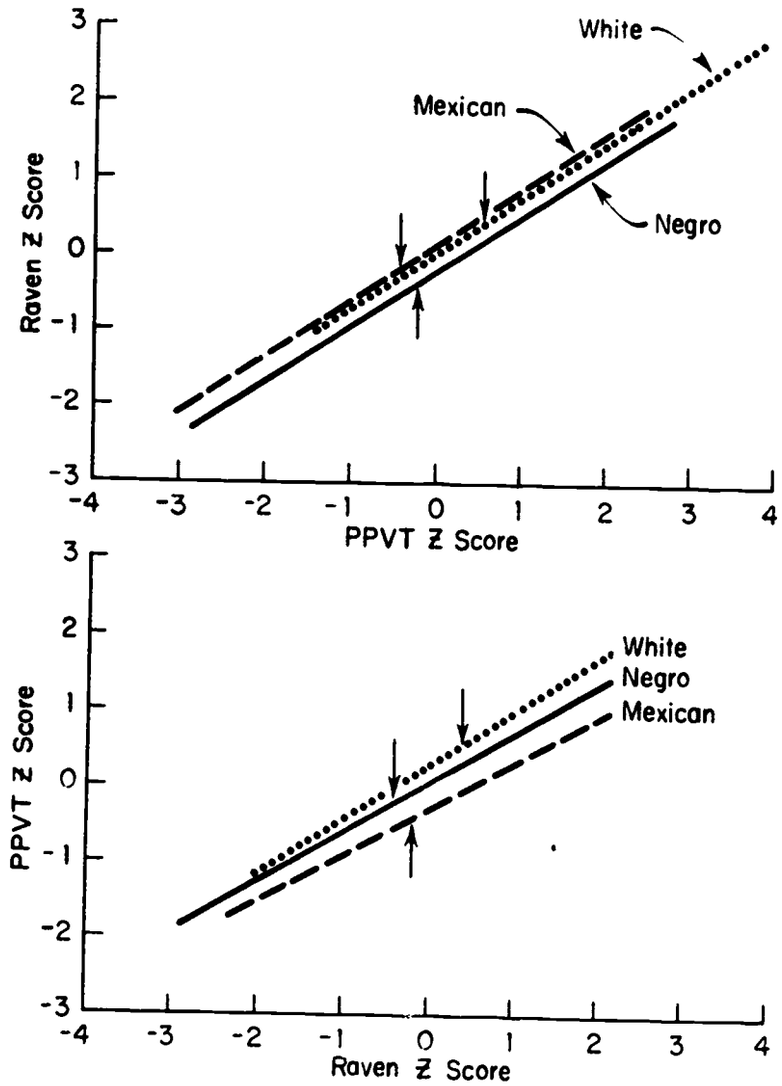

Fig. 4. Regression of Raven standardized scores (Z) on PPVT Z scores (above), and regression of PPVT on Raven (below). The bivariate means for each ethnic group are indicated by the vertical arrows.

groups obtain lower average PPVT scores than the white group. The groups' relative standing on PPVT when matched for any given Raven score is in this order, from highest to lowest: White, Negro, Mexican. But the regression of Raven on PPVT gives a quite different picture. The Mexican group accords with the culture-bias expectation, but the Negro group does not. When matched for any given PPVT score, the order of the groups on the Raven is: Mexican, White, Negro. A more complex model than the simple hypothesis that the tests merely differ in degree of culture bias favoring the majority group would seem to be necessary to explain these results. They are instructive, too, in showing that two minority groups, both socioeconomically disadvantaged relative to the white majority population, show quite different outcomes on culture-loaded and culture-reduced tests.

It may be instructive to examine how much the groups differ on the factors unique to each test. This can be shown perhaps most clearly in terms of the point biserial correlation between ethnicity and a given test score, with the other test partialed out. Since test scores show an almost perfect linear regression on age in months within two-year age intervals, the samples were divided into three approximately equal sized age groups in order to partial out age from the correlations as completely as possible prior to the main analysis. The final multiple and partial correlations are shown in Table 3. The shrunken multiple point-biserial correlation, $\underline{R}$, indexes the degree to which the various pairs of ethnic groups are discriminated jointly by the PPVT and Raven. The partial correlations are the point biserial $\underline{r}$ between the dichotomized ethnic classification (quantitized as 1 and 0) and one of the tests, with the other test partialed out.

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

Insert Table 3 about here

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

Table 3

Multiple and Partial Correlations[1] Between Test Scores and Ethnic Classification

Group	Ages 5-5 to 7-6			Ages 7-7 to 9-6			Ages 9-7 to 12-6		
		Partial r			Partial r			Partial r	
	R	PPVT	Raven	R	PPVT	Raven	R	PPVT	Raven
White (1) vs. Negro (0)	.51	.38	.18	.62	.33	.33	.61	.35	.32
White (1) vs. Mexican (0)	.63	.55	.04	.66	.44	.25	.70	.59	.14
Negro (1) vs. Mexican (0)	.29	-.28	.13	.24	-.22	.12	.36	-.32	.23

[1]Age in months partialed out of all correlations.

It can be seen that the variance unique to the PPVT and to the Raven discriminates the majority and minority groups quite differently. The PPVT and Raven discriminate whites and Negroes about equally, with the exception of the youngest age group. Much more of the discrimination between whites and Mexicans, however, is due to the PPVT; the unique Raven factor only slightly discriminates the groups. In the Negro-Mexican comparisons, the PPVT and Raven show opposite discriminations.

Reliability.--Table 4 shows the reliability of subsets of PPVT and Raven items in the three ethnic groups, determined by the Hoyt formula, which is algebraically equivalent to the Kuder-Richardson Formula 20. These reliabilities, which reflect the internal consistency of the tests, or degree of item homogeneity, are all quite substantial and reveal only negligible differences between the ethnic groups. The overall K-R reliability of the PPVT is .96 in each of the three groups. The Raven reliabilities overall are higher than the PPVT when corrected for number of items; in other words, the average item intercorrelation is higher in the Raven than in the PPVT.

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

Insert Table 4 about here

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

Item Analysis of PPVT

PPVT P Values.--The item p value is the proportion of the total sample passing the given item. The p values of the PPVT were determined for all 150 items within each ethnic groups. These are shown, averaged

Table 4

Internal Consistency Reliability[1] of PPVT and Colored Raven Matrices

	White		Negro		Mexican	
PPVT Items	Males	Females	Males	Females	Males	Females
16-30	.71	.85	.77	.66	.86	.84
31-45	.43	.88	.86	.80	.90	.88
46-60	.75	.79	.86	.84	.87	.86
61-75	.92	.92	.93	.92	.93	.91
76-90	.92	.91	.91	.87	.89	.86
91-105	.93	.94	.95	.91	.91	.93
106-120	.89	.92	.95	.94	.89	.91
121-135	.92	.93	.96	-[2]	-[2]	-[2]
All Items	.96	.96	.97	.95	.96	.95
Raven Items						
2-12	.65	.64	.58	.66	.67	.58
13-24	.79	.81	.73	.72	.80	.75
25-36	.81	.81	.70	.69	.77	.76
All Items	.90	.91	.86	.86	.90	.87

[1]Reliability determined from ANOVA using Hoyt's formula, $r_{tt} = 1 - \frac{MSV_{S\times I}}{MSV_S}$, where $MSV_{I\times S}$, the mean square variance for the Subjects × Items interaction and MSV_S is the mean square variance for Subjects.

[2]Too few Ss for a reliable estimate of r_{tt}.

over sets of 15 items, in Figure 5. The p values decrease very regularly and their rank order corresponds closely to the order of the items, which is based on the p values in the test's original normative sample in Tennessee. The three ethnic groups maintain their same relative position throughout the range of p values, though of course the discrimination is negligible at the easiest and hardest ends of the scale. It can be seen that the PPVT items comprehend a wide range of difficulty, so there is no risk of "basement" or "ceiling" effects in the ordinary school population.

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

Insert Figure 5 about here

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

One type of Race × Item interaction due to cultural differences should be reflected in differences between groups in the rank order of the individual item p values. A rank order correlation between groups of significantly less than unity, when the correlation is corrected for attenuation, is indicative of a significant Groups × Items interaction. Its magnitude is indicated by the extent of the discrepancy of the corrected correlation from 1.

Table 5 shows the rank order correlations of p values between the various ethnic groups. Since the rank order correlation of p values be-

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

Insert Table 5 about here

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

tween groups could be quite high if determined for the entire range over

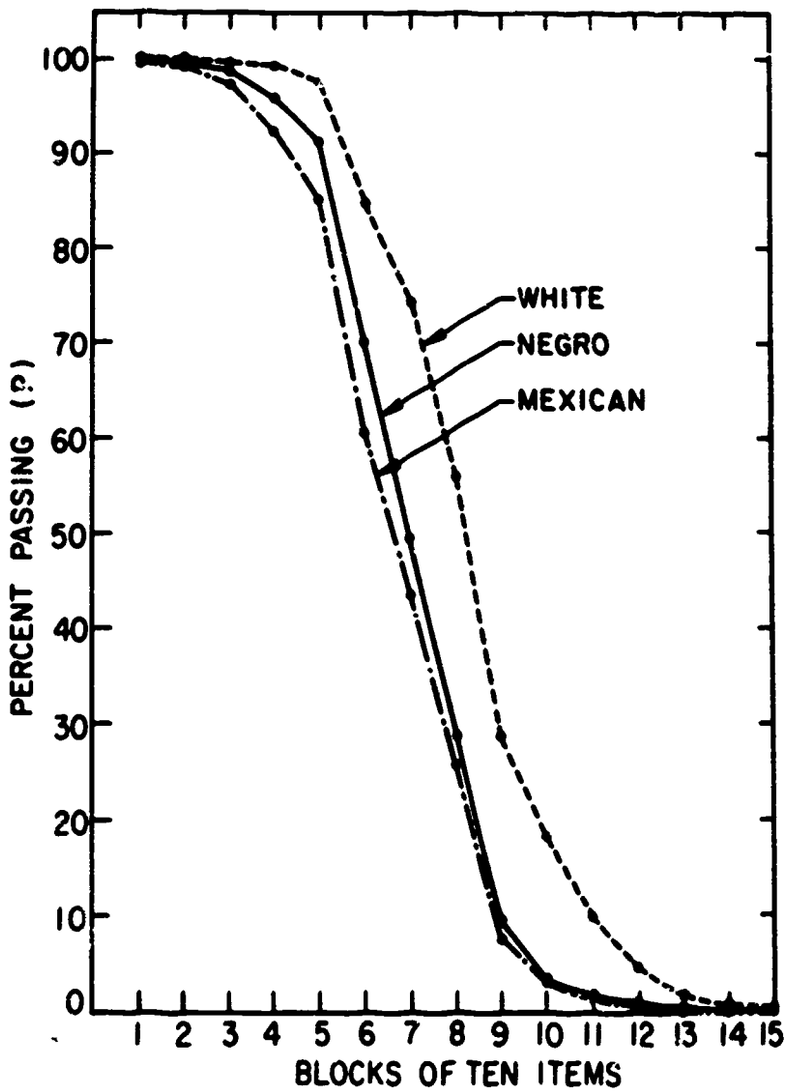

Fig. 5. Average item p values within 15-item sets of PPVT items for three ethnic groups.

Table 5

Rank Order Correlation (Corrected for Attenuation - Decimal Omitted)
Between Ethnic Groups' PPVT P Values of Items in Region of
Greatest Group Discrimination and For All Items (1-150)

Group	Items	White Female	Mexican Male	Mexican Female	Negro Male	Negro Female
White Male	31-45	688	515	492	700	538
	46-60	888	963	789	898	808
	61-75	850	963	896	896	906
	76-90	794	778	519	845	666
	1-150	988	984	978	985	980
White Female	31-45		656	582	987	1138
	46-60		945	968	911	958
	61-75		822	846	992	984
	76-90		839	807	870	819
	1-150		992	986	990	989
Mexican Male	31-45			731	842	426
	46-60			894	933	884
	61-75			958	905	905
	76-90			874	1013	949
	1-150			992	988	990
Mexican Female	31-45				800	619
	46-60				841	999
	61-75				909	901
	76-90				891	907
	1-150				982	992
Negro Male	31-45					871
	46-60					861
	61-75					977
	76-90					952
	1-150					983

all 150 items, even though the correlation may be quite low within any limited range of p values. Table 5 also shows the rank order correlations within sets of 15 items. (The first and last 15-item sets were not used because there was too little true variance to permit meaningful ranking.) The correlations are corrected for attenuation (i.e., unreliability), since we are interested in seeing if the rank order of p values is lower between groups than within groups. The reliability used in the correction for attenuation is the reliability of the rank order of p values within each of the groups being compared. These reliabilities were obtained by analysis of variance of the Items × Subjects matrix: $r_{tt}=(MSV_I-MSV_{S\times I})/(MSV_I+MSV_{S\times I})$, where MSV_I is the mean square variance for items and $MSV_{S\times I}$ is the mean square variance for the Subjects × Items interaction. These reliabilities are all extremely high, averaging close to .99, and therefore the correction for attenuation has little effect on the correlations in Table 4. But it is necessary procedure in order to determine whether the correlations remain less than 1 after correction. They obviously do, since if the true correlations were perfect, the distribution of the correlations in Table 5 should be centered about a mean of 1, with variation due to random sampling errors distributed more or less normally about the mean. It can be seen that this is not the case, so it must be concluded there is some significant degree of Ethnic Group × Item interaction in these PPVT data. However, the correlations are so high as to indicate that this form of interaction, though significant, is extremely slight, and as we shall see in a later analysis, it could be attributed to factors other than cultural differences between the groups.

The correlations are highest in those parts of the test that are the most discriminating between the ethnic groups. This is opposite to what

one should predict from a culture bias hypothesis of the group differences, which should lead to the expectation that the most discriminating items should show the least similarity between the groups in the rank order of p values. (Also note in Table 3 that the within-group reliabilities are highest in the region of the test that is most discriminating between groups.)

It is instructive to compare the correlations between ethnic groups with the correlations between sexes within each ethnic group . For the items in the four most discriminating 15-item sets (items 31-45, 46-60, 61-75, 76-90), the average correlation between the pairs of ethnic groups are:

White × Negro = .870

White × Mexican = .774

Negro × Mexican = .858

The average correlation between the sexes within ethnic groups is .861. None of these correlations differ significantly from one another. For all 150 PPVT items, the average correlation between ethnic groups is .986, and between sexes within groups is .988. In other words, a rank order of PPVT item p values differs about as little between the ethnic groups as between the sexes of the same ethnic groups.

PPVT P Decrements.--A much more sensitive index of Group × Item interaction consists of what is here called p value decrements. These consist of the difference in p values between adjacent items, e.g., 1-2, 2-3, 3-4, etc. Correlation between groups for p decrements, therefore, is not attributable to the overall regular decrease in p values from the first to the last items in all groups, but must be due to the rather slight differences in the relative difficulty of adjacent items. An indication of the sensitivity of p decrements in reflecting the relative difficulty

of items can be seen in a comparison of Forms A and B of the PPVT, consisting of entirely different stimulus words, when the two forms are correlated within a white group for p values and p decrements. The two forms were, of course, originally made up to have equal means and SDs and the items of both were arranged in the order of the p values in the normative sample. In the present study, the p values were obtained for 150 white children on Form A. These were correlated with the p values for Form B in the total white sample. The rank order correlation between the p values (over all items) of Forms A and B is .97. Yet the correlation between the p decrements of Forms A and B does not differ significantly from zero (-.014). The average correlation of p decrements between the sexes within each form, however, is .84. All this means, of course, that even if the p values are in very much the same rank order for two groups, the p decrements may not be. They reflect Group × Item interactions of the ordinal type, which do not depend upon the presence of group differences in the overall rank order of item p values, and are therefore a very subtle index of group differences in item biases.

Table 6 shows the correlation between the various groups' p decrements. These are corrected for attenuation in the same manner as described

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

Insert Table 6 about here

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

in the preceding section. The p decrements show the highest correlations in those parts of the test with the greatest between-groups discrimination. The fact that most of the corrected correlations fall below 1 indicates

Table 6

Correlation (Corrected for Attenuation - Decimal Omitted) Between Ethnic Groups' PPVT P Value Decrements in Adjacent Items in Region of Greatest Group Discrimination and For All Items (1-150)

Group	Items	White Female	Mexican Male	Mexican Female	Negro Male	Negro Female
	31-45	1200	1021	1021	1217	979
White	46-60	963	986	884	996	892
Male	61-75	905	897	796	822	825
	76-90	641	492	230	569	085
	1-150	823	778	658	786	653
	31-45		992	902	1098	1117
White	46-60		960	938	930	987
Female	61-75		916	890	933	977
	76-90		733	791	793	679
	1-150		852	809	833	874
	31-45			954	1039	778
Mexican	46-60			955	982	937
Male	61-75			978	992	913
	76-90			884	1039	808
	1-150			935	960	897
	31-45				873	905
Mexican	46-60				884	982
Female	61-75				987	931
	76-90				990	935
	1-150				873	938
	31-45					819
Negro	46-60					889
Male	61-75					959
	76-90					960
	1-150					880

a significant degree of Groups × p decrement interaction, while the magnitude of the correlations suggests that the groups are nevertheless remarkably similar in this aspect of the data. The correlations are only slightly lower than for the rank order of the p values themselves. The average correlation between the ethnic groups for the most discriminating items (Nos. 31-90) is .85; the average correlation between the sexes within ethnic groups is .93. Thus, the ethnic groups are only slightly and nonsignificantly more dissimilar than boys and girls of the same ethnic background. The fact that the correlation between the sexes is less than 1 indicates some degree of Sex × Item interaction. The overall sex difference in mean PPVT IQ, however, is negligible, unlike the ethnic group differences.

Item Analysis of Raven's Colored Matrices

Raven P Values.--For comparison of the culture loaded PPVT with a culture reduced test, the same analyses were performed on the data from Raven's Colored Progressive Matrices.

Table 7 shows the mean p values for Raven items in sets of 12 items. (Item 1 is omitted since is was used as a "practice" item while giving instructions to subjects.) The items range from easy to hard within each 12-item set, and each successive set as a whole also gradually increases in difficulty.

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

Insert Table 7 about here

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

Table 8 gives the group intercorrelations of Raven p values. These

Table 7

Mean Item P Values (Decimal Omitted) For

Raven's Colored Matrices in Three Ethnic Groups

Items	White		Negro		Mexican	
	Male	Female	Male	Female	Male	Female
2-12	782	753	663	645	709	674
13-24	675	649	503	465	555	518
25-36	554	551	381	369	409	400
All Items	667	648	511	489	553	526

are slightly higher overall than the corresponding correlations for PPVT items, indicating less Group × Item interaction, though such interaction

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

Insert Table 8 about here

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

is not completely absent since these corrected correlations are not symmetrically distributed around a mean of 1.00. The correlations between ethnic groups are:

White × Negro = .993

White × Mexican = .993

Negro × Mexican = .997,

with an overall average of .994. The average correlation between the sexes within ethnic groups is .998. In short, the ethnic groups, as well as boys and girls, are extremely alike in rank order of item difficulty in the Raven.

Raven P Decrements.--Table 9 gives the correlations between p decrements of the various groups. These correlations are nearly as high as the

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

Insert Table 9 about here

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

correlations between the rank orders of the p values, again showing a remarkable degree of similarity between the groups. The correlations between ethnic groups are:

Table 8

Correlation (Corrected for Attenuation - Decimal Omitted) Between Ethnic Groups' Colored Raven Matrices Item P Values

Group	Items	White Female	Mexican Male	Mexican Female	Negro Male	Negro Female
White Male	2-12	971	985	956	995	984
	13-24	990	986	960	964	957
	25-36	977	997	998	1000	999
	all	996	993	994	997	991
White Female	2-12		998	997	990	996
	13-24		965	981	943	936
	25-36		984	984	986	972
	all		988	997	995	990
Mexican Male	2-12			983	997	996
	13-24			985	989	987
	25-36			1004	1006	992
	all			998	999	991
Mexican Female	2-12				975	995
	13-24				970	970
	25-36				1006	992
	all				1001	996
Negro Male	2-12					997
	13-24					1005
	25-36					994
	all					1001

Table 9

Rank Order Correlation (Corrected for Attenuation - Decimal Omitted)

Between Ethnic Groups' Colored Raven Matrices

Item P Value Decrements in Adjacent Items

Group	Items	White Female	Mexican Male	Mexican Female	Negro Male	Negro Female
	2-12	1009	957	929	937	931
White	13-24	1045	1045	993	1042	852
Male	25-36	684	913	701	834	665
	all	986	993	968	982	956
	2-12		948	940	946	942
White	13-24		1008	1037	1013	965
Female	25-36		639	951	772	825
	all		977	991	984	978
	2-12			1022	1029	1005
Mexican	13-24			1006	1039	946
Male	25-36			875	1034	792
	all			991	1005	979
	2-12				1023	1014
Mexican	13-24				1014	997
Female	25-36				1017	1016
	all				1005	1001
	2-12					1035
Negro	13-24					1026
Male	25-36					1037
	all					1008

White × Negro = .982

White × Mexican = .975

Negro × Mexican = .997

with an overall average of .985. The average correlation between the sexes within ethnic groups is .995.

Correlation of PPVT Items with Ethnicity

To what degree, and how consistently, do individual PPVT items correlate with ethnicity? To find out, a measure of correlation, the phi coefficient, ϕ, which measures degree of relationship on the same scale as the Pearson r, was obtained between each item and the dichotomized ethnic variable, both for boys and girls separately and combined. The results are summarized in Table 10. The ϕ for every item was tested for

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

Insert Table 10 about here

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

significance by chi square with 1 df. It can be seen that the average Item × Ethnicity correlations are quite low, but because they are nearly all in the same direction, they add up to a considerable overall total test score × dichotomized ethnic group point-biserial correlation--about .50 for White/Negro and .60 for White/Mexican. The very few reversals of correlation, none of which are statistically significant, occur only in the later, more difficult items, which are attempted by only a small percentage of the subjects in any group. In short, there is a high level of consistency in

Table 10

Average Correlation (Phi Coefficient) of Single PPVT Items with Ethnicity

	White × Negro						White × Mexican						Negro × Mexican					
Items	Male	*	Female	*	Total	*	Male	*	Female	*	Total	*	Male	*	Female	*	Total	*
16-30	.109	2	.100	0	.113	3	.106	5	.132	1	.126	5	-.020	0	-.071	0	-.047	2
31-45	.163	7	.139	6	.152	12	.210	12	.198	12	.205	14	-.080	3	-.089	6	-.071	7
46-60	.131	10	.148	12	.140	13	.173	13	.191	13	.181	14	-.046	4	-.052	3	-.047	6
61-75	.142	10	.174	14	.155	15	.158	12	.197	12	.175	13	-.016	2	-.019	4	-.015	5
76-90	.133	10	.142	9	.136	10	.165	11	.183	9	.148	10	-.020	1	-.047	2	-.020	0
91-105	.052	1	.075	1	.056	3	.100	7	.077	2	.092	6	.036	1	-.008	1	-.021	1
106-120	-.048	1	-.002	0	-.028	2	.000	2	-.015	1	.064	4	-.096	0	-.016	0	-.018	1
121-135	.094	2	.026	0	.093	0	.108	1	-.095	0	-.008	0	-.210	0	--		+.048	0
Mean	.097		.100		.102		.127		.109		.123		-.065		-.029		-.024	
Total *		43		42		58		63		50		66		11		16		22

* Number of ϕ coefficients (within each set of 15) significant at $p < .05$.

item correlations with ethnic background. One might expect cultural biases in the strict sense to cause great discrepancies and reversals in Items × Groups correlations or discriminations, but this is not the case in the present data. It should be noted that the PPVT items were originally selected on the basis of certain psychometric properties within a white population and were not selected so as to correlate consistently with ethnic background. This property of the test is completely inadvertent. One could argue that items that correlate with ethnicity be eliminated or balanced by items that correlate in the reverse direction. Obviously, ethnically discriminating items could not be merely eliminated from the PPVT, since almost none would remain. Whether a test with otherwise similar psychometric properties could be made up that would discriminate ethnic groups in the opposite direction, yet preserve the same high degree of internal consistency reliability within all groups and the same high correlation between groups' p values and p decrements can only be determined empirically. To date no such test has been produced.

Correlations Between Raven and Special Subscales of the PPVT

Do the PPVT items which discriminate between the ethnic groups the most differ in what they measure from those that discriminate the least? To find out, special scoring keys were made up to obtain scores on subsets of PPVT items which discriminated the ethnic groups most and least, and the scores from these independent subsets of items were then intercorrelated. If the contrasting subsets actually measure different factors, their intercorrelations should be low. Moreover, if they measure the g of intelligence to different degrees, they should be expected to correlate differently with the Raven, since in factor analyses the Raven has practically all of its variance on the g factor common to a variety of measures

of mental ability. The Raven's loading on g is reported to be .80 (Raven, 1960).

To make up subtests of PPVT items that discriminate most or least between ethnic groups, the following criteria were used. The index of item discrimination was Kendall's Q, which is an index of correlation obtained from a 2 × 2 contingency table for each item (i.e., the dichotomized ethnic variable × "pass" or "fail"). Q is a monotonic function of other measures of correlation such as phi, but is on a different scale yielding a more spread-out and more normal distribution of obtained values in the present data, and mainly for this reason was used for the present analysis. Like Pearson r, Q ranges from -1 to +1. Where the cell frequencies in a 2 × 2 table are $\begin{array}{c|c} A & B \\ \hline C & D \end{array}$ $Q = (AD-BC)/(AD+BC)$. For selection of items discriminating White/Negro and White/Mexican, the least discriminating items were regarded as those with $Q < .39$; the most discriminating as those with $Q > .40$. Also, the values of $Q > .40$ had to be significant beyond $p < .05$. To insure a fair degree of reliability of the Q values, no items were used that had not been attempted by at least 100 subjects and by at least 20 subjects in whichever group of the ethnic dichotomy had the smaller number. Also, no items were used in which any of the cell frequencies in the 2 × 2 contingency table was less than 10. All the items which are useable by these criteria have positive values of Q when the ethnic dichotomies are quantitized as white = 1 and minority = 0.

The means and SDs of the Q values of the resulting subsets of the most and least discriminating items are shown in Table 11. It can be seen that

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

Insert Table 11 about here

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

Table 11

Means and SDs of Kendall's Q for the Most and the Least Ethnically Discriminating PPVT Items

Item Characteristic	Number of Items	Q Mean	Q SD
Most Discriminating:			
Whites/Negroes	33	.57	.13
Whites/Mexicans	48	.64	.16
Least Discriminating:			
Whites/Negroes	31	.24	.10
Whites/Mexicans	29	.23	.11

the subscales of the items which are the most and least correlated with ethnicity are quite separated in terms of Q.

How much do the two types of scales differ in terms of ethnic group means? Not much, it so happens, and even the least discriminating subscales show a greater mean difference between the white and minority groups than does the Raven, when all the differences are expressed in terms of sigma units, i.e., the average within-groups standard deviation. The reason is that the least discriminating subscales have smaller variances within groups as well as smaller mean raw score differences between the group means, with the result that, in terms of the average within-groups σ, the group differences are not greatly reduced by making up scales of the least ethnically discriminating items. The items that discriminate the least between groups, it turns out, are also the same items that discriminate least among individuals within the groups. Table 12 shows the mean difference (in σ units) between the white and the minority groups on the various

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

Insert Table 12 about here

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

PPVT Subscales in Grades 1 to 6. The differences on the total Raven score are given for comparison. The PPVT Subscale differences indeed come out in the expected direction, but the contrasts between the most and least discriminating subscales are surprisingly small. The contrasts, of course, would be further reduced if these scoring keys were "cross-validated" on an independent sample. It does not appear that a markedly less ethnically discriminating subscale of the PPVT can be produced by discarding the most

Table 12

Mean Difference in Sigma Units[1] Between White and Minority Groups on PPVT Subscales Consisting of the Most and the Least Ethnically Discriminating Items

PPVT Subscale	Groups	Grades K	1	2	3	4	5	6	Mean[2]
Most Discriminating:									
White/Negro	W - N	1.12	0.93	1.08	1.36	1.51	1.32	1.67	1.28
	W - M	1.52	1.38	1.56	1.74	2.00	1.67	2.31	1.71
White/Mexican	W - N	1.03	0.79	0.90	1.22	1.54	1.22	1.52	1.17
	W - M	1.58	1.40	1.48	1.82	2.22	1.82	2.21	1.79
Least Discriminating:									
White/Negro	W - N	1.06	0.88	1.11	1.03	1.02	1.17	1.25	1.07
	W - M	1.40	1.34	1.70	1.47	1.78	1.59	1.78	1.58
White/Mexican	W - N	1.09	0.96	1.21	1.13	1.14	1.42	1.49	1.21
	W - M	1.25	1.16	1.54	1.41	1.62	1.55	1.63	1.45
Raven Total	W - N	0.89	1.04	1.50	1.27	1.40	1.07	0.96	1.16
	W - M	0.61	1.09	1.42	0.92	0.95	0.97	0.67	0.95

[1]The mean difference is divided by the average σ within groups.

[2]Unweighted mean of difference (in σ units) over Grades K-6.

ethnically discriminating items. The main reason is that the items that most discriminate between the groups also most discriminate among individuals within the groups.

Do the various PPVT subscales measure different aspects or factors of ability? This is clearly not the case, since the intercorrelations among the subscales are about as high as their reliabilities will permit, and they all correlate with the Raven to much the same degree. These correlations are shown in Table 13. The most and least discriminating items

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

Insert Table 13 about here

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

appear to be measuring the same thing. If the PPVT is culture biased (as well as culture loaded) for these minorities, all the items must reflect this bias more or less uniformly. It seems remarkable indeed that from 150 culture-loaded items one cannot find a subset of items which reflect culture bias more than the rest and should therefore show a low correlation with a subset of the least biased items, and that the two subsets should correlate differently with an external criterion such as the Raven.

Equating PPVT and Raven for Difficulty

If a subset of PPVT items were perfectly equated with the Raven for difficulty in the white sample, and if the PPVT is more culturally biased against the minority groups than the Raven, one should expect a discrepancy between the white-equated PPVT and Raven scales in the minority population, with a lower mean on the PPVT than on the Raven.

Table 13

Correlations (Decimals Omitted) Between PPVT Scores Obtained with Four Different Scoring Keys and Between PPVT Scores and Raven Colored Matrices in Combined Ethnic Groups

PPVT Scoring Key	Scoring Key 1	2	3	4	Raven[1]
1. Discriminates W-N Most		91	98	89	61
2. Discriminates W-N Least			92	97	66
3. Discriminates W-M Most				88	59
4. Discriminates W-M Least					66
Number of Items in Key	33	31	48	29	

[1]Correlation of Total PPVT Score × Raven Score, in combined groups, r = .69. In white group, r = .72; Negro, r = .69; Mexican, r = .68.

To test this hypothesis, the p values of 35 items (Nos. 2-35) of Raven's colored matrices in the white male group were used as the reference. Each Raven item was matched with a PPVT item having as nearly the same p value as possible in the white group. Since there are only 35 Raven items and 150 PPVT items, it was possible with most items to achieve exact matching of p values to three decimals. In the case of exact ties, two or more PPVT items were keyed as matching a particular Raven item, and their p values were averaged in the comparison groups.

The mean p values of the matched Raven and PPVT items were then determined for all the other groups in the study. The results are summarized in Table 14. The expectations from the culture bias hypothesis show up only for the Mexican group, who perform significantly less well

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

Insert Table 14 about here

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

on the PPVT. The Negro group does not perform signficantly less well on the PPVT than on the Raven. In fact, Negro males show even slightly less difference between the PPVT and Raven than do white females. There is evidence of slightly greater though nonsignificant culture bias with respect to sex than with respect to race, as far as the White-Negro comparisons are concerned. The correlations between Raven and PPVT item p values are consistently higher for males than females, regardless of ethnicity, which further suggests a cultural sex bias in PPVT items. The last column of Table 14 shows that in more than 40 per cent of the matched pairs of items the PPVT p value exceeds the Raven p value in the Negro males.

Table 14

Summary Statistics on Raven and PPVT Scales Matched for Difficulty in the White-Male Group

Group	Mean Item P Values of Matched Scales		Raven P - PPVT P	Correlation Between Raven and PPVT P Values	Number of Matched Items on which PPVT P is Greater than Raven P
	Raven	PPVT	t Test		
White-Male[1]	.667	.667	0	1.00	--[1]
White-Female	.648	.616	0.82 n.s.	.94	16
Negro-Male	.511	.493	0.34 n.s.	.97	15
Negro-Female	.489	.408	1.62 n.s.	.93	7
Mexican Male	.553	.440	2.95*	.96	1
Mexican-Female	.526	.362	4.17*	.92	0

[1]Reference group in which the Raven and PPVT items were intentionally matched on P.

*Significant beyond .01 level.

who hardly differ from the white females in this respect. In the entire Mexican group, on the other hand, the PPVT p value exceeds the matched Raven item in only one instance. The results shown in Table 14 give some grounds for suspecting culture bias of the PPVT for the Mexican group, but not for the Negro group.

Ethnic, Sex, and Age Interactions in ANOVA

The overall most powerful means of detecting Groups × Items interactions is provided by the analysis of variance. This was applied to the present data by means of the following design: Ethnic dichotomy (2) × Sex (2) × Age (6) × Items (150 for PPVT ANOVA, 35 for Raven ANOVA), with 18 subjects per cell. The same Ss were used in both the PPVT and Raven ANOVAs. Thus there were 432 Ss in each ANOVA, with a total df = 15,119 in the Raven ANOVA and df = 64,799 in the PPVT ANOVA. In assigning Ss to the six age groups (ages 6 to 7, 7 to 8, . . . 11 to 12), Ss from each of the three ethnic groups were assigned in triplets, the members of which were matched as closely as possible for age in months, so that the means and SDs of age within each one-year interval are virtually identical in the three ethnic groups. Males and females were matched on age in the same way. Note that three ANOVAs were done for each test in order to permit pair-wise comparisons between the three ethnic groups. Putting all three groups into one ANOVA obviously would not sufficiently pinpoint the sources of variance associated with ethnicity.

Table 15 shows the complete ANOVA of the PPVT and Raven for each

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

Insert Table 15 about here

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

Table 15

Omega Squared (× 100) from ANOVA of PPVT and Raven Colored Matrices in Pairs of Ethnic Groups Matched on Age

Source of Variance	White & Negro		White & Mexican		Negro & Mexican	
	PPVT	Raven	PPVT	Raven	PPVT	Raven
Between Ss[1]	1.55**	7.47**	1.67**	7.71**	1.51**	7.24**
Ethnicity (E)	.56**	2.34**	1.18**	1.37**	.11**	.13**
Sex (S)	.10**	.11*	.05**	.03	.05**	.09**
Age (A)	1.79**	4.50**	1.62**	5.10**	1.13**	3.84**
Items (I)	73.10**	31.73**	71.47**	31.35**	76.20**	35.64**
E × S	.00	.02	.01	.00	.01	.03
E × A	.06**	.16	.09**	.30**	.01	.08
S × A	.05*	.09	.05*	.09	.06**	.09
E × I	.89**	.87**	1.50**	.47**	.21**	.20*
S × I	.21**	.19*	.14**	.27**	.16**	.20*
A × I	2.88**	1.93**	2.49**	2.74**	2.49**	2.57**
E × S × A	.02	.19	.01	.07	.02	.21
E × S × I	.07*	.11	.09**	.06	.06	.09
E × A × I	.60**	1.03*	.92**	.98*	.31	.68
S × A × I	.29	.55	.27	.54	.29	.64
E × S × A × I	.23	.67	.27	.71	.25	.51
Within Ss	17.59	48.03	18.16	48.19	17.12	47.77
Interactions:						
E and I	1.80	2.68	2.79	2.22	.83	1.47
S and I	.80	1.53	.77	1.58	.75	1.44
A and I	4.00	4.18	3.95	4.97	3.34	4.39

[1] Between Subjects within E, S, and A Groups. 18 Ss per cell.

* F for Mean Square Variance significant at $p < .05$.

** F for Mean Square Variance significant at $p < .01$.

of the possible pairs of ethnic groups. The results are presented in terms of the statistic omega squared (ω^2) × 100, which is the percentage of the total sum of squares (i.e., total variation) attributable to each source of variance. The last three rows of Table 15 show the total percent of variance attributable to all interactions (1st, 2nd, and 3rd order) involving Ethnicity × Items, Sex × Items, and Age × Items. The significance level of all the effects are indicated by asterisks. It can be seen that for all test and all ethnic group comparisons, the Ethnicity × Items interaction is significant beyond the .01 level. The more important question, however, concerns the magnitude of the interaction relative to other sources of variance.

The crucial interpretation to be drawn from Table 15 involves the magnitude of (a) the Ethnicity main effect relative to the Subjects (within groups) main effect, and (b) the Ethnicity × Items interaction relative to the within-group Subjects × Items interaction. The extent to which the test discriminates between the ethnic groups, relative to the discrimination between subjects within groups, is indicated by the ratio of the main effect for Ethnicity to the main effect for Subjects (within groups). The extent to which items are biased (i.e., show interaction) with respect to ethnic groups relative to the interaction of Items × Ss within groups is indicated by the ratio of the interaction of Ethnicity × Items to the interaction of Ss (within groups) × Items. We are forced to compare the variances in terms of ratios, since the ethnic group differences are interpretable only in relation to individual differences within groups. Ideally, in a culture-reduced test the ratio of main effects (i.e., Ethnicity/Ss)should be large relative to the ratio of interactions (i.e., Ethnicity × Items/Ss × Items). A large Ethnicity × Items interaction relative to the Subjects × Items

interaction would mean that some particular selection of items from the same population of items that compose the test could be found that would have satisfactory reliability and could equalize or reverse the mean scores of the two ethnic groups. A very small Ethnicity × Items interaction relative to the Ss × Items interaction tends to rule out this possibility. It would mean that no subset of items could be found with satisfactory reliability which would equalize or reverse the ethnic group means.

Table 16 shows these ratios, and the last two columns, A/B, shows

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

Insert Table 16 about here

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

their relative magnitudes for the PPVT and the Raven.[4] (Ignore the last row of Table 16 until reading the next section.) The main effects ratios are much greater than the interaction ratios, which is what should be expected of tests with little ethnic group bias, as here defined. As indicated in the A/B columns of Table 16, the Raven shows considerably less of the "undesirable" interaction than the PPVT in discriminating the white and minority groups. By this criterion, however, even the PPVT shows very little item bias. Also, by the same criterion, the tests show greater sex bias than ethnic bias. The A/B ratio for sex (averaged over the 3 sets of comparisons in Table 15) is 4.24 for the PPVT and 2.37 for the Raven. With A/B ratios this small, careful item selection could stand a chance of equalizing or reversing the slight sex difference on these tests. All this, of course, is highly consistent with the previous analyses in terms of the high correlations between the ethnic groups in p values and p decrements.

Table 16

Variance Ratios for (A) Ethnic Main Effect/Subjects Main Effect and (B) Ethnic × Item Interaction/Subjects × Item Interaction and the Ratio A/B, for PPVT and Raven Tests in Various Ethnic Comparisons

Groups	(A) Main Effects Ratio		(B) Interaction Ratio		A / B	
	PPVT	Raven	PPVT	Raven	PPVT	Raven
White and Negro	.361	.313	.051	.018	7.10	17.32
White and Mexican	.706	.178	.083	.010	8.55	18.13
Negro and Mexican	.075	.018	.012	.004	6.07	4.46
White Older and Younger	.473	.347	.059	.019	7.97	18.26

Age × Item Interaction.--So far, therefore, it appears that there is a statistically significant but very small degree of test bias as indicated by the item interactions with ethnicity. But now notice in Table 15 that there is also a considerable Age × Item interaction. This raises the question of whether the ethnic group differences and item interactions reflect not cultural differences, but merely the same kinds of differences and item interactions that result from differences in mental maturity, as reflected by age-group differences, within any ethnic group.

Can the ethnic effects shown in Table 15 be simulated by making up "pseudo-ethnic" groups composed of younger and older children within any one ethnic group? To find out, two "pseudo-ethnic" groups were formed as follows: one group consists of 96 younger white Ss between the ages 6 and 9 (assigned to three age groups in one-year intervals); the other group consists of 96 older white Ss between the ages 8 and 11 (assigned to three age groups in one-year intervals). Note that the younger and older groups overlap in age, but they have a mean age difference of two years. The two groups composed by this particular selection according to age were called "pseudo-ethnic" groups because the chronological age differentials within and between the two groups were made to approximate, as closely as feasibly possible, the average mental age differential between the white and Negro groups in the total sample. In other words, by means of age selection, two white groups were composed that would simulate the means and variances of the total white and Negro populations, respectively. The two all-white pseudo-ethnic groups were formed strictly by age selection; Ss were not included or excluded in terms of their individual performance on the tests.

The item data for PPVT and Raven of these two pseudo-ethnic groups, labeled Older and Younger were subjected to the same ANOVA (except there

were three instead of six age groups) as was used with the real ethnic groups shown in Table 15. The results of the ANOVA for the "pseudo-ethnic" groups are shown in the last two columns of Table 17. Compare these percentages

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

Insert Table 17 about here

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

of variance for all the various main effects and interactions with those for the white and Negro ANOVA shown in the first two columns of Table 15. There is hardly any difference! And the true ethnic and "pseudo-ethnic" main effects and interactions differ least of all. In short, the same evidence of ethnic "culture" bias can be produced within a culturally homogeneous sample simply by selection of two different chronological age groups which differ in mental age to about the same extent as the mental age difference between whites and Negroes when these groups are matched on chronological age. This means that the magnitude of Group × Item interactions that are seen in Table 15 are not at all dependent upon ethnic cultural differences but can occur in a culturally homogeneous population strictly as a result of differences in mental maturity. Returning to Table 16, the last row permits comparison of the ratios for the true ethnic groups and the pseudo-ethnic groups (i.e., white Older and Younger). Note the great similarity to the true white and minority results, especially in the critical A/B ratio.

If the ethnic group effects can thus be simulated within a culturally homogeneous sample, the question arises, can the Ethnicity × Item interaction be appreciably reduced in an ANOVA which compares younger whites with older ethnic group children, with the chronological age differential made such as

Table 17

Omega Squared (× 100) from ANOVA on PPVT and Colored Raven Matrices Given to White Children (Ages 6-9) and Minority Children (Ages 8-11) and to Two Groups of White Children--Younger (Ages 6-9) and Older (Ages 8-11)

Source of Variance	White (Ages 6-9) and Negro (Ages 8-11)		White (Ages 6-9) and Mexican (Ages 8-11)		White Younger (Ages 6-9) And Older (Ages 8-11)	
	PPVT	Raven	PPVT	Raven	PPVT	Raven
Between Ss[1]	1.50	7.88	1.54	7.60	1.59	7.45
Ethnicity (E)	.00	.14	.11	.00	.75	2.58
Sex (S)	.09	.19	.04	.02	.07	.10
Age (A)	.64	2.49	.63	3.63	1.15	2.87
Items (I)	78.12	35.60	77.39	34.15	73.34	30.16
E × S	.02	.05	.00	.01	.02	.00
E × A	.02	.34	.01	.06	.00	.49
S × A	.05	.16	.03	.09	.01	.04
E × I	.12	.22	.30	.26	1.10	.94
S × I	.24	.33	.17	.48	.25	.43
A × I	1.36	1.37	1.56	2.01	1.93	1.67
E × S × A	.01	.08	.01	.08	.01	.11
E × S × I	.09	.19	.10	.11	.12	.17
E × A × I	.22	.59	.24	.55	.55	1.09
S × A × I	.24	.65	.24	.54	.21	.73
E × S × A × I	.18	.61	.19	.59	.23	.71
Within Ss	17.10	49.12	17.43	49.81	18.64	49.45
Interactions:						
E and I	.61	1.61	.83	1.52	2.01	2.91
S and I	.75	1.77	.69	1.72	.81	2.04
A and I	2.00	3.21	2.23	3.70	2.93	4.19

[1] Between Subjects Within E, S and A Groups. The ANOVAs in the first 4 columns have 18 Ss per cell. ANOVAs in the last 2 columns have 16 Ss per cell.

to minimize the mean mental age difference between the ethnic groups entering into the ANOVA? To accomplish this, whites of ages 6 to 9 were compared with minorities of ages 8 to 11 in the same ANOVA as before. The results are shown in the first two pairs of columns in Table 17. The main effect of Ethnicity is practically eliminated, as was intended, but why should the Ethnicity × Items interaction be so greatly reduced (e.g., by 87% in the white and Negro ANOVA) if it reflects culture bias? The cultural backgrounds of the groups under comparison have not been changed in the least, but only their ages. If one argues that cultural handicap is overcome increasingly with age, then we should expect there to be a regular convergence of white and minority scores going from younger to older age groups. As can be seen in Figures 1, 2, and 6, this is not the case.

The results of all these ANOVAs in which age was manipulated are more consistent with a hypothesis of differences in mental maturity interacting with items than of ethnic cultural differences producing such interaction. The main effect of ethnicity is subject to the same interpretation, unless one posits that ethnic cultural factors should have a more or less uniformly depressing effect on all 150 items of the PPVT and on all 35 items of the Raven.

Study II. PPVT and Raven in Socioeconomically Extreme White and Negro Groups

The Ss in the preceding study were a representative cross section of all children in a California school district in which there are not very extreme socioeconomic contrasts within or between the ethnic groups. Study II, on the other hand, examines the PPVT and Raven's Colored Progressive Matrices in perhaps the most extremely contrasting neighborhood schools

with respect to SES background to be found in a California school district (Contra Costa County). The population of Contra Costa encompasses as extreme socioeconomic diversity as is likely to be found in any California school district.

The two schools from which the present samples were randomly drawn were all white and all Negro.[5] The former is located in an upper-middle class suburb, the latter in a low SES Negro neighborhood. The neighborhoods were specifically selected from census tract information on the basis of such SES indices as median income, median educational level, percentage of homeowners, average value of dwellings, average rent, ratio of deteriorating and dilapidated dwellings to "sound" dwellings, and a crowding index. The white and Negro groups are widely separated and totally non-overlapping on all these indices. The modal occupational category of the "head of household" as entered in the school records was "professional" or "managerial" in the white school and "unskilled" or "welfare" in the Negro school. The two schools differ at least 30 points in average IQ. The contrasting groups are clearly not typical of the general white and Negro populations. But these greatly contrasting groups are highly appropriate for the present study. Whatever is the nature of the cultural differences making for test biases that are claimed to exist between the general white and Negro populations, such culture biases should only be exaggerated in the white and Negro groups selected for the present study.

Subjects.--24 Ss of each sex were selected at random from each of Grades K, 1, 3 in the white and Negro schools, making the total N = 288. The average age of the white sample was 6 yr. 11 mos.; of the Negro sample, 7 yr. 2 mos.

Tests.--The PPVT and Raven's Colored Progressive Matrices were administered individually in two one-hour sessions, separated by 2 to 5 days. The PPVT was given according to the standard directions given in the test manual (Dunn, 1965). The presentation of the Raven was preceded by four similar practice problems which aided in making clear the instructions; these practice problems were presented like a form board so that the S could easily get the idea of how one particular pattern from among the multiple-choice alternatives would complete the total matrix pattern when it was inserted into the blank space in the matrix formboard. All Ss were encouraged to attempt all 36 items of the Raven. The fact that the average percent passing the first 4 items of the Raven test proper was 98.4% for the white group and 98.4% for the Negro group is a good indication that the Ss of both groups clearly understood the instructions and requirements of the test.

Results

Mean Group Differences.--The average differences expressed in average white-group σ units between the white and Negro groups are shown in Table 18. The white-Negro differences are very similar on both the PPVT and the Raven

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

Insert Table 18 about here

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

with the exception of the kindergarten group, in which there is a much smaller difference on the Raven. At the higher grade levels, however, the groups differ on the Raven at least as much as they differ on the PPVT.

Table 18

Mean Differences in σ Units Between White and Negro Groups at Three Grade Levels on PPVT and Raven's Colored Matrices

Grade	PPVT	Raven
K	1.69	0.54
1	1.31	1.32
3	2.42	2.46

P Values and P Decrements.--Table 19 shows the item p values averaged within sets of items and the correlations between the white and Negro p values and p decrements within these sets of items. The correlations (not corrected for attenuation) are remarkably high, especially for the Raven. The very substantial correlations for the p decrements is also noteworthy, considering the sensitivity of this index in reflecting differences in the difficulty of adjacent items.

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

Insert Table 19 about here

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

PPVT and Raven Matched for Item Difficulty.--As in the previous study, PPVT items in the white group were matched as closely as possible with all 36 Raven items on the basis of item p values. The correlation between the white-matched p values of PPVT and Raven p values was 1.00 for the white group and .95 for the Negro group. The mean Raven and PPVT values for Negroes were .417 and .348, respectively, which, though in the expected direction, is not a significant difference even at the .10 level.

The procedure was also reversed, i.e., the Raven and PPVT item p values were matched in the Negro sample, with a correlation of 1.00. Their correlation in the white group was .87. If the PPVT is more culturally biased than the Raven in favor of upper-middle-class whites, we should expect the white sample to obtain a higher mean p on the PPVT than on the Raven. In fact, the mean p values of the white group on the Negro-matched Raven and PPVT were .575 and .613, respectively; again in the expected direction, but nonsignificant ($t < 1$). In short, even in these extremely contrasting race and SES groups, the PPVT does not appear markedly more culture-biased than the Raven. The magnitude of the difference between the matched PPVT and

Table 19

Mean P Values (Decimals Omitted) for Whites and Negroes Within Subsets of PPVT and Raven Colored Matrices, and Correlations[1] Between White and Negro P Values and P Decrements

PPVT Items	Mean P Value White (N = 144)	Mean P Value Negro (N = 144)	Correlation Between P Values	Correlation Between P Decrements
31-45	982	873	.77	.71
46-60	794	610	.86	.88
61-75	489	169	.76	.80
Mean	755	551	.80	.80
Raven Items				
1-12	703	589	.95	.87
13-24	565	379	.88	.69
25-36	475	283	.94	.73
Mean	575	417	.92	.76

[1]Not corrected for attenuation.

Raven within each group (when the matching was done on the other group) is trivial compared to the magnitude of the difference between the racial-SES groups on either test. These results are inconsistent with a hypothesis of culture bias or verbal deprivation affecting the culture-loaded vocabulary test appreciably more than the nonverbal culture-reduced test. If cultural differences or deprivations exist in the low SES Negro group as compared with the upper-middle SES white group, these results indicate that the cultural bias must more or less uniformly depress performance on both types of test items as well as on all the items within each type of test.

Analysis of Multiple-Choice Distractors.--When white and Negro children make errors on the PPVT, do they make different errors? Is there some kind of cultural difference that would prompt the white and Negro children to choose different distractors when they are not sure of the correct response? Every PPVT item has one possible correct response and three distractors. A chi square analysis was performed on every set of distractors to determine if the relative frequency of choices differed in the white and Negro groups. Only those items were used which were attempted and missed by at least 15 Ss in each racial group, in order to insure adequate sensitivity of the chi square test for detecting a significant association between choice of distractor and racial group. There were 23 PPVT items which qualified for this analysis. Of the 23 chi square tests, six (or 26%) were significant beyond the .05 level. This is obviously greater than chance. When the total sample was randomly divided in half and the chi square test was performed in each half, the same six items showed a significant racial difference in choice of distractor. (These were items 48, 52, 59, 61, 70, 71.) But oddly enough, the white and Negro p values of these particular items do not differ more, on the average, than the white and Negro p values of other

items on which the two groups do not differ in the choice of distractors. The question arises, are these merely differences in sheer guessing tendency on certain items? If there was pure guessing, the proportion of responses to each of the three distractors should be quite equally divided among them, close to 1/3 for each. The size of the standard deviation of the proportions on each distractor should therefore be an index of departure from random guessing. Whites showed a larger SD on three and Negroes showed a larger SD on three of the six sets of distractors that showed significant chi squares. So there does not appear to be any consistent evidence of a racial-SES difference in guessing tendency.

The same kind of chi square analysis of distractor choice was performed on Raven items 5 to 36. Four of the 32 items (11, 12, 29, 32) showed racial group differences in the choice of distractor significant beyond the .05 level. The white-Negro p values on these particular items do not differ more than for other items, which, as in the PPVT, means that whatever biases determine the choice of distractor are not necessarily the same as those that affect the difficulty of the item.

Most Popular Response.--There are four possible alternative responses (including the correct response) to each PPVT item. Is the most popular response alternative (i.e., the response selected by the largest percentage of Ss) different in the white and Negro samples? This was examined for all PPVT items which were attempted by at least 40 Ss in each racial group. Only 6 of the 71 items of the PPVT showed the most popular response to be different in the two groups, and these all cross-validated when the groups were randomly divided in half. Usually, of course, the most popular response in both groups was the correct response.

The 36 Raven items showed no ethnic group differences at all in the

most popular response to each item, even when the most popular response was one of the erroneous distractors.

From the analysis of distractors and most popular responses, it appears that the Raven shows less signs of race-SES bias than the PPVT, though whatever bias is reflected by these indices seems unrelated to race differences in item difficulty per se.

Study III. Analysis of Raven's Matrices in Three Ethnic Groups

In order to look more closely at the developmental lag hypothesis of test differences and also to detect possible ethnic biases in Raven's Matrices multiple-choice distractors over a much wider range of ages than was possible in the previous study, the following analyses were performed on large representative samples of three ethnic groups in Grades 3 to 8 who had been given the Colored Matrices (Grades 3 to 6) and the Standard Matrices (Grades 7 and 8).

Subjects and Tests

Ss were representative samples of children from a large school district in the Central Valley (Kern County) of California. Raven's Colored Matrices was group-administered to regular classes in Grades 3 to 6, with approximately equal numbers in each grade. The three ethnic groups are white (N = 841), Negro (N = 687), and Mexican-American (N = 788).

The Standard Progressive Matrices, which consists of 60 items and extends from very easy items up to a level of difficulty appropriate for the general adult population, was group-administered to classes in Grades 7 and 8. The Ns are white = 744, Negro = 551, and Mexican-American = 608.

Results

Descriptive Statistics.--Figure 6 shows the performance of the three ethnic groups at each grade level in terms of t scores with an overall mean of 50 and SD of 10 (based on the SD of raw scores in the white group at Grade 5). (It was possible to put the Standard Matrices given to Grades 7 and 8 on the same scale as the Colored Matrices given in Grades 3 to 6, since for other purposes both tests were given to subsamples ranging from Grades 4 to 8 so the standardized scores of the two tests could be made continuous over the entire grade range.)

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

Insert Figure 6 about here

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

P Values and P Decrements.--Table 20 shows the mean p values and ethnic group correlations between p values and between p decrements for 12-item sets of the Colored Matrices (Grades 3 - 6). Table 21 shows the corresponding results for the Standard Matrices (Grades 7 and 8).

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

Insert Tables 20 and 21 about here

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

The rank order of the three ethnic groups' p values on each item are highly consistent, with W > M > N. In fact, only three of the 60 items of the Standard Matrices depart from the order W > M > N, and they are very dfficult items (36, 58, 60) which less than 8% of any group answered

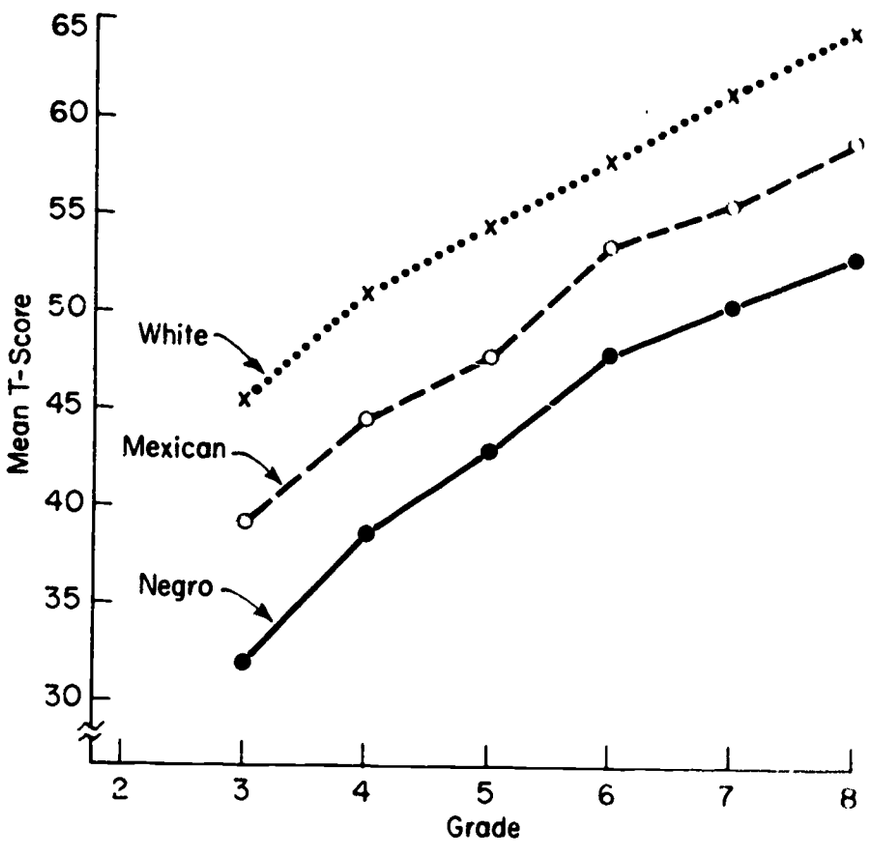

Fig. 6. Mean T scores (mean = 50, SD = 10) on Raven's Progressive Matrices.

Table 20

Correlations[1] Between Ethnic[2] Groups' Colored Progressive Matrices

Item P Values and P Decrements, and Mean P Values

Items	Correlation for P Values			Correlation for P Decrements			Mean P Values		
	W × N	W × M	N × M	W × N	W × M	N × M	W	N	M
1-12	.96	.99	.98	.79	.94	.95	.827	.723	.775
13-24	.96	.99	.97	.88	.91	.86	.762	.598	.685
25-36	.96	.98	.99	.73	.77	.94	.650	.460	.547
Mean	.96	.99	.98	.80	.87	.92	.746	.594	.669

[1]Not corrected for attenuation

[2]White (W), Negro (N), Mexican-American (M).

Table 21

Correlations[1] Between Ethnic[2] Groups' Standard Progressive Matrices Item P Values and P Value Decrements, and Mean P Values

Items	Correlation for P Values			Correlation for P Decrements			Mean P Values		
	W × N	W × M	N × M	W × N	W × M	N × M	W	N	M
1-12	.96	.99	.99	.81	.94	.94	.905	.825	.871
13-24	.93	.94	.99	.82	.83	.87	.776	.605	.670
25-36	.95	.99	.99	.82	.95	.88	.612	.447	.516
37-48	.97	.99	.99	.88	.97	.93	.645	.465	.558
49-60	.96	.99	.98	.77	.85	.93	.282	.161	.206
Mean	.95	.98	.99	.82	.91	.91	.644	.500	.564

[1]Not corrected for attenuation.

[2]White (W), Negro (N), Mexican-American (M).

correctly and on which the ethnic groups do not differ significantly.

The correlations between ethnic groups in item p values could hardly be higher. The within-group reliabilities of the rank order of item p values are of about the same magnitude. Note also the size of the correlations for the p decrements. These results give every indication that both forms of Raven's Matrices behave extremely alike in all three ethnic groups. If there are cultural differences, they are surely not revealed by this type of analysis.

When the p value correlations are determined in each grade separately, it turns out that whites resemble Negroes who are about 2 years older, more than they resemble Negroes of the same age or other whites who are two years older. (The same thing is not true in comparing whites and Mexicans on the Raven.) For example, Grade 4 whites are more like Grade 6 Negroes (r = .978) than Grade 4 whites are like Grade 6 whites (r = .806). This result seems less consistent with the hypothesis of a cultural difference than with a difference in rate of mental development, unless it is assumed that test manifestations of cultural differences are indistinguishable from the test manifestations of general developmental differences.

Cultural Differences vs. Developmental Lag.--To examine this notion more closely, Raven's Colored Matrices items were subjected to a principal components analysis separately in each ethnic group in each of Grades 4, 5, and 6. Interest is focused on the first principal component, which of course accounts for the largest proportion of item variance and indicates the loading (i.e., correlation) of each item on the general factor of mental ability which is common to all the items in the test. In a sense, the items' loadings on the first principal component represent a weighting of the items from which has been screened out that part of the variance contributed by

factors that are unique to each item or which only certain subsets of items share in common. The loadings would therefore seem less likely to reflect differential cultural biases than the unweighted item scores of 0 or 1.

The question of main interest here involves the degree of resemblance in the first principal component between different grades (i.e., age groups) within ethnic categories as compared to the resemblance between the ethnic groups (both within and across grades). Degree of resemblance is determined by the correlation between groups' item loadings on the first principal component. The rank order correlation was used, so that there would be equal means and variances of the variables entering into each correlation, permitting direct comparisons of the obtained correlations. In each ethnic group in each grade, the g loadings (i.e., first principal components) of the 36 Raven items were ranked from 1 to 36, and the rank order correlations between all possible Grades × Ethnic Groups were obtained. These correlations are shown in Table 22.

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

Insert Table 22 about here

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

The pattern of intercorrelations is of primary interest. We see, for example, that on this measure Grade 4 whites resemble Grade 5 whites less than Grade 6 Negroes, although Grade 4 whites resemble Grade 4 Mexicans more than Mexicans in any other grade. In general, resemblance across grades within ethnic groups (mean rho = .46) is slightly less than resemblance between ethnic groups (mean rho = .50), and in the case of the white-Negro comparisons, resemblance is greatest between whites and Negroes who are

Table 22

Rank Order Correlation[1] Between Grades (4, 5, and 6) and Ethnic Groups on Loadings of First Principal Component for Raven's Colored Matrices Items

		White			Negro			Mexican		
Group	Grade	4	5	6	4	5	6	4	5	6
	4		.67	.14	.65	.59	.85	.75	.28	-.02
White	5			.12	.54	.59	.71	.59	.31	.28
	6				.56	.51	.33	.43	.68	.27
	4					.73	.77	.67	.56	.31
Negro	5						.71	.68	.68	.18
	6							.75	.51	.18
	4								.49	.14
Mexican	5									.37
	6									

[1]All correlations larger than 0.50 are significant beyond .01.

separated by one or two grade levels. This is summarized in Table 23, in

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

Insert Table 23 about here

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

which the correlations between the pairs of ethnic groups are averaged over (a) those in the same grade, (b) those separated by one grade, where the Negro group is always the higher grade, and (c) those separated by two grades, where the Negro group is always the higher grade. Note that the white × Negro correlation increases with amount of grade separation, and the Negro × Mexican correlations are parallel in this respect. But the White × Mexican correlations go in the opposite direction and the resemblance is greatest between the groups in the same grade. Thus, according to this analysis, the Negro group appears to fall more in line with the hypothesis of a developmental lag rather than of a cultural difference. The Mexican group, on the other hand, does not accord with expectations from the developmental lag hypothesis in this analysis.

Analysis of Distractors.--A chi square test was performed on the frequencies of choice of the five error distractors for each of the 36 Colored Matrices items to determine if there were any significant differences between the ethnic groups in the choice of distractors. The entire sample of 2,316 Ss was used.

Four items showed differences in choice of distractors significant at the .05 level. This is above the chance expectation. On three of the items (23, 31, 36) the significant difference in distractor choice was between whites and Negroes, with the largest percentage difference on any

Table 23

Average Correlation (Rho) Between Ethnic Groups' First Principal Component Loadings on Raven's Colored Matrices Items When Groups Are in Same Grade or Are Separated by One or Two Grades

Averaged Correlations	Correlated Ethnic Groups[1] W × N	W × M	N × M
Same Grade	.52	.44	.51
Separated 1 Grade[2]	.65	.28	.60
Separated 2 Grades[2]	.85	-.02	.75
Mean	.67	.23	.62

[1]White (W), Negro (N), Mexican-American (M).

[2]Negro grade is always higher.

of the distractors being 15%, 12%, and 11% respectively. One item (3) had a significant Negro-Mexican difference of 16% on the most discriminating distractor. On none of these items is the minority group's p value significantly or consistently less than for other items which have similar p values in the white group but show no significant ethnic difference in the choise of distractors.

The same kind of analysis was performed on the 60 items of the Standard Progressive Matrices given in Grades 7 and 8, with a total N = 1,903. Four items (19, 35, 47, 50) showed significant (.05 level) white-Negro differences of 19%, 17%, 10% and 19% for the most discriminating distractors. One of the same items (35) also showed a significant white-Mexican difference of 22%. The minority groups do not have lower p values on these items than on others of the same approximate difficulty in the white sample.

Most Popular Response.--Do the ethnic groups differ in their selection of the one out of six multiple-choice alternatives (including the correct one) that they choose most frequently? In the 36 Colored Matrices items, six were found in which a different response alternative was more "popular" for one ethnic group than for the others and which also cross-validated in two random halves of the total sample. The items on which this occurred tended to be the most difficult ones for all three ethnic groups (12, 24, 32, 33, 35, 36) and therefore they would have relatively little overall effect on the group means. This can be shown by making up several special scoring keys, each based on the most popular responses in a given ethnic group being keyed as "correct." If cultural biases lead to systematically different solutions to matrix items, then one might argue that different scoring keys might be more appropriate for different groups. So three scoring keys

based on the most popular responses in the white, Negro, and Mexican groups were made up in one random half of each sample and "cross-validated" on the other random half. Every key was applied to every group. It turns out that no matter which scoring key is used, the ethnic group means are consistently in the order W > M > N, and the differences between the means are in every case significant beyond the .01 level.

Of the 60 Standard Progressive Matrices items, only one (53) showed an ethnic difference (W - N) in the most popular response alternative which cross-validated in two random halves of the total sample. Thus different ethnic scoring would involve only one item in the Negro group, and since it is one of the most difficult and least discriminating items in all three groups neither the elimination nor the re-keying of the item would make virtually any difference in the average Raven scores of the three groups.

Even if different ethnic scoring keys were found which equalized or reversed the orders of the group mean scores, it still would have to be determined if such scoring keys also reduced the mean score differences between ethnically and culturally homogeneous age groups separated by one or two years. It is likely that the choice of particular distractors in preference to others is more related to an individual's degree of mental maturity than to his cultural background per se. One indication of this is seen in the fact that on the five items of the Colored Matrices which showed a difference between whites and Negroes in the most popular response alternative chosen, there is a greater similarity in the choice of distractors between younger whites and older Negroes than between whites and Negroes of the same age, and the difference between younger whites and older whites resembles in this respect the difference between Negroes and whites of the same age. This is shown in Table 24. These figures were obtained as follows:

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

Insert Table 24 about here

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

The items were those on which the most popular response alternative for Negroes (total sample) was different from the most popular response alternative for whites (total sample). (In all five items, the most popular response both in the Negro and in the white groups is an "incorrect" response according to the standard scoring key.) Among all those who failed the given item was determined the percentage of Negroes and whites in combined Grades 3 & 4 and in combined Grades 5 & 6 who chose the distractor which is most popular for Negroes (total sample). The relevant differences between these percentages are the figures shown in Table 24. Note that the difference between Grade 5 & 6 Negroes and Grade 5 & 6 whites in choice of the most popular Negro distractor is considerably greater than the difference between Grade 5 & 6 Negroes and Grade 3 & 4 whites, who average about two years younger in age. Moreover, the difference between Grade 3 & 4 whites and Grade 5 & 6 whites more closely resembles the difference between the Negro and white groups in the same grade (i.e., 5 & 6). In other words, the distractors most commonly chosen by Negroes of a given age are also the same distractors that are more frequently chosen by whites who average about two years younger. Thus the tendency to be "taken in" by a particular distractor appears to be more a function of the Ss mental age than of his racial-cultural background per se.

Summary and Discussion

An important distinction is made between culture-loaded and culture-biased, as these terms are applied to mental tests. Culture loading is

Table 24

Difference in Percentage[1] of Negroes and Whites in Grades 3 & 4 and Grades 5 & 6 Who Chose the Distractor Most Often Chosen by the Total Negro Sample on the Five Raven Colored Matrices Items for Which the Most Popular Response Alternatives Differed for Negroes and Whites

Item Number	Distractor Number	(Negro % in Grades 5 & 6) - (White % in Grades 5 & 6)	(Negro % in Grades 5 & 6) - (White % in Grades 3 & 4)	(White % in Grades 3 & 4) - (White % in Grades 5 & 6)
24	4	3.30	.90	2.40
32	5	9.00	1.70	7.30
33	1	-5.10	-3.00	-2.10
35	2/3[2]	10.40	2.00	8.40
36	2	11.00	5.60	5.40
Mean		5.72	1.44	4.28

[1] Percentage based only on the total number of Ss in each group who failed the item, i.e., who chose one of the five distractors.

[2] Distractors 2 and 3 had the same percentage in total Negro group and so the percentages in this analysis are the average of the two.

defined in terms of types of item content and the narrowness of the cultural background to which the content of the test items is relevant or is likely to be encountered by members of different subpopulations. Culture bias is defined in terms of various external and internal criteria. External criteria involve the test's predictive validity in different ethnic or cultural groups, as assessed by the regression of measurements of some external criterion (e.g., grades, job performance ratings, etc.) on test scores. Internal criteria involve item characteristics which may vary statistically between different cultural groups, such as differences in the rank order of item difficulties, groups × items interaction, group differences in choice of distractors for items answered incorrectly, group differences in reliability, item intercorrelations, and factor loadings of test items.

Group mean differences per se are not evidence of bias, since the causes of the goup differences may be essentially the same as the causes of individual differences within the groups. The notion of culture bias implies that the cause of a group mean difference is qualitatively different from the cause of individual differences within groups.

The presence of a substantial groups × items interaction is presumptive evidence of culture bias unless the interaction can be equally well accounted for by some counter hypothesis. The absence of a substantial or significant groups × items interaction in the presence of a significant groups main effect, however, cannot prove that the group mean difference is not due to some cultural or environment factor, if it is hypothesized that the factor influences all of the test items about equally. The plausibility of such a hypothesis would depend largely upon the nautre of the hypothesized factor. It would seem more plausible, for example, that malnutrition or poor motivation would have a generalized effect on performance

which would quite equally depress performance on a wide variety of test items. Cultural group differences, on the other hand, would seem more likely to have differential effects on various items or types of test content, thereby producing a marked groups × items interaction, or groups × type-of-test interaction, such as between verbal and nonverbal tests. Those who claim that tests are biased either against or in favor of one or another ethnic or cultural group are obligated to produce evidence that such bias in fact exists in terms of some objective set of criteria, external or internal. Culture-loaded test content or group mean differences do not by themselves constitute evidence of bias with reference to the particular groups in question. Test bias relates to particular groups. It is not a property of the test itself.

In the present series of studies, a highly culture-loaded test, the Peabody Picture Vocabulary Test, and a culture-reduced test, Raven's Progressive Matrices, were examined for internal evidence of culture bias in comparisons between large representative samples of white, Negro, and Mexican-American children from three California school districts.

The main findings are as follows:

The internal consistency reliability (Kuder-Richardson Formula 20) is very high and practically the same in the white, Negro, and Mexican-American samples, for both the PPVT and the Raven Matrices. When corrected for differences in length (i.e., number of items), the Raven has slightly higher K-R reliability than the PPVT.

Both the Raven and PPVT show similar correlations with chronological age (in months) for all three ethnic groups, although the correlations on both tests are highest for whites. This may be due in small part to the fact that in the white group test scores show a slightly more linear regression on age than in the two other groups, where there are slight departures

from linear regression. But the lower correlations with age in the minority groups are attributable mostly to the fact that in these groups the regression of raw scores on age has a less steep slope than in the white group, i.e., the average year-to-year gains are smaller in the minority groups (particulary for the Mexicans on the PPVT and for the Negroes on the Raven), and this fact, along with the nearly equal standard deviations of raw scores in all three groups, makes for slightly lower correlations with age in the minority groups. An essential characteristic of intelligence tests in the age range from early childhood to maturity is that the raw scores correlate with chronological age. This means, of course, that individual differences in test scores at any given age represent much the same kinds of differences in degree of mental maturity typically observed between younger and older children. One criterion of the validity of newly devised tests intended to minimize the effects of cultural bias is the demonstration of a correlation with age in the target population comparable to the age correlations found for existing standard tests in their normative population.

When the groups are compared in the rank order of item p values (percent passing an item), they are found to be highly similar, as indicated by very high rank order correlations between the item p values in all three groups, correlations which, when corrected for attenuation, are very close to 1, even when the correlations are computed within subsets of 12 or 15 items (for Raven and PPVT, respectively). By this criterion neither test shows much evidence of culture bias, as would be indicated by dissimilarities in the rank order of item difficulty in the various ethnic groups. As expected, the Raven items show somewhat higher ethnic group similarities in relative difficulty than the PPVT.

The differences between adjacent items in percent passing (called p

decrements) are highly similar in all three ethnic groups for both PPVT and Raven. The intergroup similarity in this sensitive index indicates little of the groups × items interaction that should be expected if the test items were ethnically biased to varying degrees. The high degree of similarity between the ethnic groups in p decrements suggests that the groups behave very much the same on these tests except for mean differences in the total number right.

When PPVT and Raven are exactly matched item-by-item for difficulty in the white group, and the matched scales are then compared in the Negro and Mexican groups, the Negro group showed no difference in means on the white-matched PPVT and Raven scales, while the Mexican group showed a significantly lower mean on the PPVT than on the Raven. This indicates that the Mexican group is somewhat handicapped on the culture-loaded PPVT relative to the culture-reduced Raven, but the Negro group is not. The fact that the Mexican group is very similar to the white in rank order of p values and p decrements on both PPVT and Raven, yet has lower scores on the PPVT than on the Raven, suggests that some factor is operating to depress the PPVT performance more or less uniformly for all items and that this factor does not depress Raven performance, at least to the same degree. It seems plausible to suggest that this factor is verbal and may be associated with bilingualism in the Mexican group. The Negro group does not show this discrepancy between performance on the PPVT and the Raven; the Negro performance deficit is about the same on both tests, as different as they are in culture loading.

Correlations (phi coefficient) of single PPVT items with the ethnic dichotomy white/minority are all positive when signficant; no PPVT items discriminate significantly in the reverse direction. When separate PPVT

scales are made up, consisting either of the least or of the most ethnically discriminating items, the ethnic group mean differences are not markedly different on the two scales when measured in sigma units, since the standard deviations are less in these specially derived subscales. The items that discriminate most between the ethnic groups are also the same items that discriminate most among individuals within each group. This finding is the opposite to what should be expected if the test from which these subscales were derived was highly culture biased. Moreover, there is no evidence that the least and most discriminating subscales measure different factors, since their intercorrelation is about as high as reliability permits.

The ethnic groups differ more than chance in the most frequent choice of item distractors in the PPVT and Raven. However, on the few Raven items on which the most popular response choice differs for whites and Negroes, it turns out that the most popular distractor for Negroes is the same as the most popular distractor for whites who are approximately two years younger. This suggests that the choice of particular distractors in the Raven is related to S's mental maturity. If total score on the test reflects differences in mental maturity (as indicated by the substantial correlation of raw scores with age in all three groups), and if the choice of distractors is related to mental maturity, then groups that differ in mean total score might be expected to show some differences in their modal choice of distractors, and the types of group differences should be similar to the differences seen between younger and older Ss within groups. If the choice of distractors were influenced mainly by cultural differences, they would be less likely to coincide with the distractor choices that are related to age differences within a culturally homogeneous group.

In other findings, also, ethnic group differences in average cognitive

maturity seems a more parsimonious explanation than culture bias, especially in the case of the Negro samples. For example, the matrix of Raven item intercorrelations within each ethnic group within each grade, from Grades 3 to 6, was subjected to a principal components analysis. The loadings of items on the first principal component (the general factor common to all Raven items) were compared between age groups and ethnic groups. Degree of group similarity was measured by the correlation of the loadings of the 36 Raven items in each of the two groups being compared. Within the same grade, resemblance is higher between whites and Mexicans than between whites and Negroes. But the resemblance between whites and Negroes was greater for groups separated by two grade levels. Negroes in Grade 6, for example, were more similar to whites in Grade 4 than to Negroes in Grades 4 or 5. The Mexican group, on the other hand, showed their greatest resemblance to whites in the same grade.

Analysis of variance of the complete Groups × Items × Subjects matrix provides the most sensitive and powerful means for detecting internal evidence of culture biases in test items. This ANOVA was performed on the same randomly selected Ss from each of the three ethnic groups for both the PPVT and Raven. For this analysis the ethnic groups were matched for age. The three ANOVAs involved each of the possible group comparisons--White/Negro, White/Mexican, and Negro/Mexican. Sex and age were also included as factors in the ANOVA. For both the PPVT and the Raven, the interaction of ethnic group × items was significant, although it accounts for an exceedingly small proportion of the total variance. The crucial index of culture-fairness, however, is the ratio of the sum of squares of the (A) Between Ethnic Groups Main Effect/Between Subjects Within Groups Main Effect to the (B)Groups × Items Interaction/Ss × Items interaction. Lower values of this A/B ratio

indicate item biases with respect to groups, and higher values indicate less item bias. The higher the A/B ratio, the more difficult it should be to equalize or reverse the group mean difference by item selection from the same general population of items of which those comprising the particular test are a sample. It is noteworthy, therefore, that the A/B ratio for the culture-reduced Raven is more than double that for the PPVT. Also, for the PPVT, a higher ratio (i.e., less item bias) is found in the White/Negro than in the White/Mexican comparison. The A/B ratio can be applied as well to sex differences, using the appropriate main effects and interactions. Sex shows item biases of even greater magnitude than the ethnic biases and the Raven shows less sex bias than the PPVT. The very low A/B ratio for sex, especially on the PPVT, suggests that a different selection of similar items, or even merely discarding some of the existing items, could eliminate or reverse the small sex difference in means and it may therefore be regarded as a trivial or nonessential difference. The same thing cannot be said, however, about the mean ethnic group differences, for which the A/B ratio is probably much too great to permit elimination of the group differences by any amount of item selection from the item pool constituting the PPVT and the Raven. One wonders if any set of items could be found to form a test which would reverse the group means and still preserve all of the other desirable psychometric characteristics seen in the PPVT and the Raven. As of the present time, there has yet been no such demonstration.

The Groups × Items interaction can be all but eliminated if the ANOVA is based on a white group and a minority group which differ about one or two years in average age. Then the younger white group and older minority group have nearly equal total mean scores and the Groups × Items interaction is practically nil, both for the PPVT and the Raven. In other words, the

small Groups × Items interactions found in the same-age ethnic group comparisons can be interpreted as reflecting a mental maturity × items interaction rather than a cultural difference × items interaction. It would seem far-fetched to argue that the Groups × Items interaction reflects culture bias when such interaction can be greatly reduced simply by comparing ethnic groups that differ one or two years in age. If it is argued that the effect of culture bias on test performance decreases as children get older, then one should also find a decrease in the mean difference between ethnic groups with increasing age. Yet the mean differences are at least as great, absolutely and in standard deviation units, in older as in younger age groups.

The hypothesis that the ethnic groups × items interaction reflects differences in mental maturity more than culture bias is reinforced by the fact that it was possible to simulate almost exactly the results of the White/Negro ANOVA by making up a "pseudo-ethnic" group of whites. In this, two white groups were compared, using the same ANOVA design as in the true ethnic group comparisons. One of the white groups was selected so as to average two years older than the other white group. The two age groups (both white) took the place of the two ethnic groups in the ANOVA. The main effects and the Groups × Items interaction almost exactly simulated the White/Negro ANOVA; and the A/B ratios, of course, were also nearly the same. This was true both for the PPVT and the Raven. This finding suggests the conclusion that little or none of the Group × Items interaction in the case of the Negro samples is attributable to cultural differences.

The evidence regarding cultural or language bias in the Mexican group is less clear. Some of the findings are consistent with the hypothesis that in the Mexican group PPVT performance is depressed, relative to the Raven.

In the rank order of the three ethnic group means, Negroes and Mexicans reverse positions on the Raven and PPVT. When white and Mexican Ss are matched for PPVT scores, the Mexicans have a higher mean score on the Raven, which is what is to be expected if the PPVT performance was depressed by some factor peculiar to a culture-loaded test but not to a culture-reduced test. On the other hand, when white and Negro Ss are matched for PPVT scores, the Negroes also have a lower mean score on the Raven, and this holds throughout the entire range of scores.

Viewed all together, the present set of analyses reveal very little, if any, evidence of culture bias in either the PPVT and Raven for the Negro group. Also, the Raven shows practically no evidence of bias in the Mexican group. However, the extent of bias in the PPVT with respect to the Mexican group is more in doubt; the evidence for bias is not strong but it is not ruled out by the present analyses, some of which are consistent with the predictions from a culture bias hypothesis. But without exception, this is not true for the Negro group. If culture bias is claimed for the Negroes, it must also be posited that the bias affects all items of the PPVT and the Raven about equally. This seems most improbable for a cultural effect. It is more likely attributable to other factors that could be reasonably hypothesized to have a much more general influence on overall rate of mental development.

If it is claimed that the ethnic group differences in average performance on tests such as the PPVT and Raven are mainly the result of cultural differences, then it should be possible to make up other tests which are biased in favor of the ethnic minority groups, and yet at the same time show the same psychometric properties as the present tests, such as a small Groups × Items interaction, a large A/B ratio, high intergroup correlations

between p values and between p decrements, as well as similar correlations with age in each group and equally high internal consistency reliability within the different groups. The construction of a test that could equalize or reverse the white and Negro group means, and which also could stand up under the kinds of analysis to which the PPVT and Raven were subjected in the present studies, would be a strong challenge to any theory which holds that the average racial difference in IQ is not attributable to cultural bias in the tests.

References

Buros, O. I. (Ed.) Sixth Mental Measurements Yearbook. Highland Park, New Jersey: Gryphon Press, 1965. Pp. 820-823.

Cleary, T. A. Test bias: Prediction of grades of Negro and white students in integrated colleges. Journal of Educational Measurement, 1968. 5, 115-124.

Cleary, T. A., & Hilton, T. L. An investigation of item bias. Educational and Psychological Measurement, 1968, 28, 61-75.

Council of the Society for the Psychological Study of Social Issues. Statement by SPSSI on current IQ controversy: Heredity versus environment. American Psychologist, 1969, 24, 1039-1040.

Darlington, R. B. Another look at "cultural fairness." Journal of Educational Measurement, 1971, 8, 71-82.

Dunn, L. M. Expanded Manual, Peabody Picture Vocabulary Test. Minneapolis: American Guidance Service, Inc., 1965.

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. Another look at culture-fair tests. In Western Regional Conference on Testing Problems, Proceedings for 1968, "Measurement for Educational Planning." Berkeley, Calif.: Educational Testing Service, Western Office, 1968. Pp. 50-104.

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

MacArthur, R. S., & Elley, W. B. The reduction of socioeconomic bias in intelligence testing. British Journal of Educational Psychology, 1963, 33, 107-119.

Mercer, Jane R. & Brown, W. C. Racial differences in IQ: Factor artifact. In Senna, C. (Ed.) The fallacy of I.Q. New York: The Third Press, 1973. Pp. 56-113.

Raven, J. C. Guide to the Standard Progressive Matrices. London: H. K. Lewis, 1960.

Stanley, J. C. Plotting ANOVA interactions for ease of visual interpretation. Educational and Psychological Measurement, 1969, 29, 793-797.

Thorndike, E. L., & Lorge, I. The teacher's word book of 30,000 words. New York: Teachers College Press, 1944.

Thorndike, R. L. Concepts of culture-fairness. Journal of Educational Measurement, 1971, 8, 63-70.

Williams, R. L. Abuses and misuses in testing black children. Counseling Psychologist, 1971, 2, 62-77.

Footnotes

[1] Much of the data collection and analyses in the present studies were supported by grants to the University of California from the Office of Economic Opportunity (Contract No. OEO 2404) and the Sterling Morton Charitable Trust.

[2] The writer is grateful to Dr. Mabel C. Purl, Director of Research and Evaluation, Riverside Unified Schools, for these data.

[3] Table of the p values for each item of the PPVT and Raven within each sex and ethnic group is available from the author.

[4] Note that the same A/B ratio can also be obtained (from Table 15) by $\frac{E}{E\times I} \Big/ \frac{Ss}{Ss\times I}$. The A/B ratio can also be expressed as $F_E \Big/ F_{E\times I}$, where F_E is the variance ratio for testing the significance of the Ethnic main effect and $F_{E\times I}$ is the variance ratio for testing the Ethnicity × Items interaction.

[5] The writer is grateful to Dr. William D. Rohwer, Jr. for these data.