JOURNAL OF EDUCATIONAL MEASUREMENT
VOLUME 11, NO. 1 SPRING 1974

THE EFFECT OF RACE OF EXAMINER ON THE MENTAL TEST SCORES OF WHITE AND BLACK PUPILS[1]

ARTHUR R. JENSEN
University of California, Berkeley

An entire elementary school system with 60% white and 40% black pupils was given several abiity tests group-administered by 12 white and eight black examiners (*E*s). The tests measured verbal and nonverbal IQ, perceptual-motor cognitive development, "speed and persistence" under neutral and motivating instructions, listening-attention, and short-term rote memory for numbers. With the exception of the "speed and persistence" test, on which white *E*s yielded significantly and consistently higher mean scores than black *E*s for both white and black pupils across grades one to six, the results for the various cognitive ability tests showed that the race of the *E* did not produce large or consistent effects in the testing of white and black pupils.

How often is it said that the race of the examiner is an important variable in the ability testing of ethnic minority children, or that black children obtain lower scores when tested by a white examiner? Sattler (1970, p. 144), in a review of the research on this question, remarked:

> In spite of the paucity of research concerning the effects of differences in racial status as a variable which affects the examiner-examinee relationship, numerous writers have either concluded or suggested that this variable may play an important role in the intelligence test situation.

And in this and another review (Sattler & Theye, 1967, p. 353) Sattler cites a dozen references holding this belief, including books and articles by such noted psychologists as Anastasi, Hilgard, Klineberg, Pettigrew, Pressey, and Strong.

Though the speculative claims of race of examiner (*E*) effects in intelligence testing are frequent in the literature of race differences, the total empirical research on the subject, 11 studies and a reanalysis of one of these, altogether constitute a rather unimpressive body of evidence. They are briefly summarized chronologically here:

> Canady (1936): On the first administration of 1916 Stanford-Binet, *S*s obtained higher IQ with *E*s (one black and 20 whites) of their own race, while on retest *S*s obtained higher IQs with *E*s of the opposite race. Sattler (1966) reanalyzed the experiment and concluded the results are inconclusive because of methodological deficiencies.
>
> Pasamanick and Knoblock (1955): A white *E* testing 40, two-year-old black *S*s on the Gesell Developmental Examination was claimed to have obtained lower verbal responsiveness scores than presumably would have been obtained by a black *E*, but no *S*s were tested by a black *E* for comparison.

[1]The collection of these data was funded by a grant from the Berkeley Unified School District to the University of California. Statistical analysis of the data was made possible by a grant from the Sterling Morton Charitable Trust.

The writer is especially indebted to Dr. Wade Egbert for scheduling and supervising the testing and the scoring of the tests, to Dr. Jane B. Brooks for assistance in training the testers, and to Dr. Carol Treanor for the computer processing of the data.

Forrester and Klaus (1964): Twenty-four black kindergarteners obtained nonsignificantly higher Stanford-Binet IQs when tested by a female black *E* than by a female white *E*.

La Crosse (1964): A white *E* obtained significantly lower Stanford-Binet (L-M) retest scores when testing black *S*s who had been previously tested by two black *E*s. The same white *E* obtained significantly higher retest scores with white *S*s previously tested by three white *E*s.

Pettigrew (1964): White *E*s (number not reported) are said to obtain fewer correct responses than black *E*s (number not reported) from northern blacks given two tests (identification of six famous men and giving synonyms). No statistical tests of significance are reported.

Miller and Phillips (1966): Three black and three white female *E*s testing black and white children in Head Start in the south resulted in no significant effects, either for race of *E*, or for the race of *E* × race of *S* interaction.

Pelosi (1968): Six black and six white *E*s tested young adult black males enrolled in a Neighborhood Youth Corps, on the Wechsler Adult Intelligence Scale, the Purdue pegboard, and the IPAT Culture Fair Test; no significant effects of race of *E*.

Abramson (1969): Two black and two white female *E*s gave Peabody Picture Vocabulary Test to eastern black and white kindergarteners and first-graders. No significant *E* effects for kindergarteners, but white *E*s obtained higher scores from white *S*s than from black; and black *E*s obtained similar scores from both groups.

Lipsitz (1969): Lorge-Thorndike group-administered test-retest by one black and one white *E* showed no significant race of *E* or interaction effects in eastern black and white fourth, fifth, and sixth graders in private schools (unrepresentative samples).

Caldwell and Knight (1970): Stanford-Binet test-retest with one black female *E* and one white male *E* produced no significant *E* effect on sixth grade southern black children.

Costello (1970): Two white and two black *E*s giving the Peabody Picture Vocabulary Test to black preschoolers resulted in no significant race of *E* effect.

Six of the above studies show no significant race of *E* effects; two studies yield a significant effect, at least in one or another grade or group; and three studies involve no statistical test or are wholly inconclusive because of methodological shortcomings. In a detailed critique of this body of research, Sattler (1973) commented:

> The studies reviewed . . . suggest that performance of Negro and white subjects on individually administered intelligence tests is not usually affected by the examiner's race. However, there are still too few studies available to arrive at firm generalizations. Yet, as Sattler (1970, p. 144) pointed out, numerous authorities have stated that difference in racial status is a variable which affects the examiner-examinee relationship. The research cited in this review as well as past research offers no support for this statement.

Sattler (1970, p. 144) also points out that "little is known about the effects of the examiners' race on scores obtained on group administered intelligence tests." To examine this matter in terms of available evidence, Shuey (1966) compared all the reported studies up to 1965 (19 in all) of black IQ in elementary school children in the south, where the group testing was done by a black tester, with the test results obtained on all southern black school children, the vast majority of whom were tested by white examiners. Shuey concluded (Ibid., p. 507):

> The 2,360 elementary school children tested by Negroes earned a mean IQ of 80.9 as compared with a combined mean of 80.6 earned by more than 30,000 southern Negro school children, an undetermined but probably a large number of whom were tested by white investigators. The present writer also calculated the combined mean IQ achieved by 1,796 southern colored high school pupils who were tested by Negro adults. This was 82.9 as compared with a mean of 82.1 secured by nearly 9,000 southern colored high school students, many of whom were examined by white researchers. From these comparisons it would seem that the intelligence score of a Negro school child or high school pupil has not been adversely affected by the presence of a white tester.

The most obvious source of possible bias in Shuey's analysis is the fact that there was no control over the samples tested by black and white *E*s. If for some reason the less intelligent black *S*s were more likely to be tested by black *E*s (as might be the case in most rural southern schools), the fact that the more intelligent black *S*s (more likely in urban schools) tested by white *E*s did not obtain higher IQs than the *S*s tested by black *E*s might only mean that their performance had been depressed by the presence of a white *E*. In a proper study pains should be taken to avoid any such biasing factors in the assignment of white and black *E*s to white and black *S*s.

The present study was designed to control such factors. It was conducted with large enough samples of both *E*s and *S*s over a sufficient age range and with an adequate variety of mental ability tests as fully to permit the significant appearance of a race of *E* × race of *S* interaction. Since statistical significance depends in part upon the sample size, and since the samples in this study are very large, it is more important to evaluate the actual magnitudes of the examiner effects rather than merely to note their level of statistical significance.

METHOD

Subjects

The *S*s were virtually the total white and black elementary school (kindergarten through sixth grade) population of the Berkeley Unified School District. A total of nearly 9,000 pupils in all classes of 17 schools were tested, with the exclusion only of children in special classes for the retarded, the emotionally disturbed, and the neurologically and physically handicapped. Since the present study focuses upon white-black interaction of race of examiners (*E*) and race of subjects (*S*s), the 11 percent of the school population who are Oriental or other ethnic minorities (about 1%) are not included in the analyses. (The total school population involved in this study is approximately 60% white and 40% black.) Also not included are *S*s who were absent on the day that a particular test was administered to their class.

*S*s' ethnicity was determined from the school records, which included the parents' statement of the child's race, obtained when the child was enrolled in the Berkeley schools.

Examiners

There were 12 white (ten women and two men) and eight black (six women and two men) *E*s. All were between 25 and 40 years of age and all had either B.A. or M.A. degrees in psychology or education. A few were university graduate students in the school psychology program and nearly all of them had teaching credentials and had taught in public schools. They were paid at the daily rate for substitute teachers in the Berkeley schools.

All *E*s were given copies of the tests and the manuals of instructions for administration to study prior to the three all-day training sessions. These sessions, conducted by three professionals with training and experience in clinical and group testing, aimed to inculcate general principles of test administration as well as specific instructions and practice in the tests to be used. All testing procedures were demonstrated and all *E*s, working in small groups, had to practice the instructions and procedures in front of the

group and the instructor who criticized and "shaped up" each *E*'s performance in terms of voice, emphasis, pacing, rapport, and general manner of presentation. The importance of strict adherence to the standard instructions and time limits was emphasized repeatedly and the psychometric rationale for this was thoroughly explained. *E*s were fully aware that one of the main purposes of this training was to minimize, as much as possible, examiner differences as a source of variance in test scores.

*E*s were provided with stopwatches for the timed tests and were taught to operate the tape recorders used in two of the tests. *E*s were also instructed in filling out a special form at the conclusion of every test session concerning any unusual occurrences (e.g., a fire drill) which might have created nontypical testing conditions.

All *E*s were observed actually testing in the classroom at least once, early in the testing program, by Dr. Egbert, our testing supervisor, or one of the other professionals on the staff, with the aim of maintaining as much uniformity of testing procedures as possible.

All *E*s did not administer every one of the different tests used in this study, but every test was administered by white and black *E*s, and by male and female *E*s.

Assignment of Es *to Schools and Classes*

The assignment of *E*s to schools and classes was random within race of *E*. That is, on any given day, one black *E* was assigned at random to each school until the supply of black *E*s was used up; the same was done for white *E*s. Thus, every school received both white and black *E*s. These random assignments were made on a day-to-day basis, so that all *E*s had equal chances of testing in all schools. The particular classes to be tested at a given school on a given day also were assigned at random to the white and black *E*s. Because of the unequal numbers of female and male *E*s and the relatively small number of the latter, no attempt was made formally to include the sex of *E* as part of the analysis. The random assignment of *E*s, as described above, was applied without regard to sex, so that the numbers of white and black pupils tested by male and female *E*s is roughly proportional to the relative frequencies of male and female *E*s of each race.

Tests

A variety of quite different tests were used. They were expected possibly to elicit different degrees of sensitivity to examiner effects. There were standard verbal and nonverbal IQ tests which involved considerable verbal instructions on the part of *E*, especially in grades kindergarten to three. There was an untimed developmental perceptual-motor test; a "speed and persistence" test intended to reflect effort and motivation induced by verbal instructions in a test-taking situation; a test of *S*s' ability to attend to verbally given directions; and a short-term memory test. Both of these latter two tests involved the presence and supervision of the *E* as a proctor, but were wholly administered and paced by means of a tape recording to insure perfect uniformity of instructions, pacing, and the like.

Lorge-Thorndike Intelligence Tests. This is a nationally standardized group-administered test of general intelligence (Buros, 1959, pp. 478–484).

The tests for grades kindergarten to three do not depend at all upon reading ability but make use exclusively of pictorial items. The tests for grades four to eight consist of two parts, Verbal (V) and Nonverbal (NV).

The following forms of the Lorge-Thorndike Intelligence Tests were used.

Level 1, Form B.	Grades K-1
Level 2, Form B.	Grades 2-3
Level 3, Form B, Verbal and Nonverbal.	Grades 4-6.

The "consumable" form of the test was used to obviate separate answer sheets and the added difficulty they may involve for the testees.

Figure Copying Test. This test was developed at the Gesell Institute of Child Study at Yale University as a means for measuring developmental readiness for the traditional school learning tasks of the primary grades (Ilg & Ames, 1967). The test consists of the ten geometric forms arranged in order of difficulty. The child must simply copy them, each on a separate sheet of paper. The test involves no memory factor since the figure to be copied is before the child at all times. The test is administered without time limit, although most children finish in 10 to 15 minutes. The test is best regarded as a developmental scale of mental ability. It correlates substantially with other IQ tests, but it is considerably less culture-loaded than most usual IQ tests.

Each of the ten figures is scored on a 3-point scale going from 1 (low) to 3 (high). (A score of zero is given in the rare instance when no attempt has been made to copy a particular figure.)

Listening-Attention Test. In the Listening-Attention Test the child is presented with an answer sheet containing 100 pairs of digits in sets of ten. The child listens to a tape recording which speaks one digit every 2 seconds. (The recorded male voice is very clear and lacks any distinctive regional accent, being similar in quality to that of most network radio and television announcers.) The child is required to put an X over the one digit in each pair which has been heard on the tape recorder. The purpose of this test is to determine the extent to which the child is able to pay attention to numbers spoken on a tape recorder, to keep his place in the test, and to make the appropriate responses to what he hears from moment to moment. High scores (95% or more correct) on the Listening-Attention Test indicate that the subject possesses the necessary subskills for taking the digit span (Memory for Numbers) test. Low scores (less than 90% correct) show up pupils who, for whatever reason, are unable to hear and to respond to numbers read over a tape recorder, and for whom, therefore, the Memory for Numbers Test is probably not a valid measure of short-term memory. The Listening-Attention Test itself makes no demands on the child's memory, but only on his ability for listening, paying attention, and responding appropriately—all prerequisites for the digit memory test that follows.

Memory for Numbers Test. The Memory for Numbers test is a measure of digit span, or more generally, short-term memory. It consists of three parts. Each part consists of six series of digits going from four digits in a series up to nine digits in a series. The digit series are presented on a tape recording on which the digits are spoken clearly by a male voice (the same as in the Listening-Attention Test) at the rate of precisely one digit per second. The subjects write down as many digits as they can recall at the conclusion of each series, which is signaled by a "bong." Each part of the test is preceded by a short practice test of three digit series in order to permit the tester to determine whether the child has understood the instructions, etc. The practice test also serves to familiarize the subject with the procedure of each of the subtests. The first subtest is la-

beled Immediate Recall (I). Here the subject is instructed to recall the series *immediately* after the last digit has been spoken on the tape recorder. The second subtest consists of Delayed Recall (D). Here the subject is instructed not to write down his response until after 10 seconds have elapsed after the last digit has been spoken. The 10-second interval is terminated by the sound of a "bong" which signals the child to write his response. The Delayed Recall condition invariably results in some retention decrement. The third subtest is the repeated series test in which the digit series is repeated three times prior to recall; the subject then recalls the series immediately after the last digit in the series has been presented. Again, recall is signaled by a "bong." Each repetition of the series is separated by a tone with a duration of 1 second. The repeated series almost invariably results in greater recall than the single series.

Speed and Persistence Test (Making Xs). The Making Xs Test is intended as an assessment of test-taking motivation. It gives an indication of the subject's willingness to comply with instructions in a group testing situation and to mobilize effort in following those instructions for a brief period of time. The test involves no intellectual component, although for very young children it probably involves some perceptual-motor skills component. Individual differences among children at any one grade level would seem to reflect mainly general motivation and test-taking attitudes in a group situation. Most children without a motor handicap who do very poorly on this test, it can be suspected, are likely not to put out their maximum effort on ability tests given in a group situation.

The Making Xs Test consists of two parts. On Part I the subject is asked simply to make Xs in a series of squares for a period of 90 seconds. In this part the instructions say nothing about speed. They merely instruct the child to make Xs. The maximum possible score on Part I is 150, since there are 150 squares provided in which the child can make Xs. After a 2-minute rest period the child turns the page of the test booklet to part II. Here the child instructed to show how much better he can perform than he did on Part I and to work as rapidly as possible. The child is again given 90 seconds to make as many Xs as he can in the 150 boxes provided. The gain in score from Part I to Part II reflects both a practice effect and an increase in motivation or effort as a result of the motivating instructions.

RESULTS AND DISCUSSION

The basic analysis performed on each test at each grade level in which the test was administered is a nested ANOVA, with race of *E*s nested within race of *S*s. Since there were unequal *N*s in the four cells of the 2 × 2 design, the main effects for race of *E* are based on unweighted means for white and black *E*s; that is to say, in the overall means for white and black *E*s, equal weights are given to both means despite their unequal *N*s. Otherwise the overall mean difference between white and black *E*s would be partly a function of the number of white and black *S*s they had tested, because there is a substantial main effect for race of *S*s.

So that the magnitudes of the differences can be readily compared from one grade to another and from one test to another, all differences have been expressed in sigma units. The sigma in every case is the standard deviation of test scores within groups, i.e., the standard deviation excluding variance due to race of *E*, race of *S*s, and their interaction.

Table 1 shows the *N*s of *E*s and *S*s for each of the tests at each grade.

TABLE 1

Number of Examiners (in Italics) and Number of Subjects (in Roman type) of Each Race (W = White, B = Black) in Each Grade for Each Test

Grade		K		1		2		3		4		5		6		Total *S*s	
Tests	*E*s	W	B	W	B	W	B	W	B	W	B	W	B	W	B	W	B
Lorge-Thorndike Nonverbal IQ	W	*8* 468	*8* 333	*7* 447	*6* 294	*7* 394	*6* 350	*5* 308	*5* 247	*7* 361	*7* 248	*7* 298	*7* 184	*8* 366	*8* 237	2642	1893
	B	*3* 132	*3* 58	*3* 194	*2* 107	*4* 297	*4* 139	*3* 273	*2* 148	*5* 226	*4* 165	*6* 254	*6* 205	*4* 248	*4* 166	1624	988
Lorge-Thorndike Verbal IQ	W									*6* 363	*7* 270	*8* 393	*7* 89	*8* 434	*7* 208	1190	567
	B									*5* 202	*6* 100	*4* 161	*5* 302	*4* 171	*5* 151	534	553
Figure copying	W	*8* 291	*8* 268	*9* 363	*9* 289	*11* 510	*11* 337	*11* 386	*11* 222	*8* 226	*8* 211					1776	1327
	B	*7* 340	*6* 211	*7* 321	*7* 196	*6* 174	*6* 153	*7* 239	*7* 157	*7* 257	*7* 207					1331	924
Speed and Persistence	W			*11* 507	*11* 325	*10* 440	*10* 309	*10* 456	*10* 261	*9* 378	*10* 281	*11* 429	*12* 308	*10* 446	*10* 245	2656	1729
	B			*7* 195	*8* 172	*6* 169	*7* 187	*5* 157	*5* 150	*6* 247	*7* 150	*7* 167	*7* 111	*6* 179	*6* 144	1114	914
Listening-Attention and Memory	W					*4* 471	*4* 347	*4* 358	*4* 260	*4* 330	*4* 300	*5* 337	*5* 289	*5* 374	*5* 270	1870	1466
	B					*2* 178	*2* 114	*2* 219	*2* 133	*2* 221	*2* 85	*2* 210	*2* 103	*2* 197	*2* 99	1025	534

Lorge-Thorndike IQ

Tables 2 and 3 show the results for Lorge-Thorndike Nonverbal and Verbal IQs, respectively. For Nonverbal IQ, the main effect of race of *E*, as we can see in the first column, is very small; only at grades one and two is the difference significant, and it amounts to less than one-fifth of a standard deviation, or less than three IQ points. (Negative numbers always indicate that the black mean exceeded the white.) The significant effect of race of *E* in grades one and two on Nonverbal IQ is attributable to the white *S*s scoring significantly higher under white *E*s and black *S*s scoring higher under black *E*s. The interpretation of these small but statistically significant differences is made problematic by the fact that they occur only in certain grades, are not consistent in direction from one grade to another, nor for Nonverbal and Verbal IQ, and do not appear to follow any consistent trend across grades. Despite their statistical significance (and quite small effects can be significant with such large *N*s), these effects do not appear to be very systematic or psychologically interpretable in the present context.

TABLE 2
Lorge-Thorndike Nonverbal IQ: Mean Differences in Sigma Units

Grade	Mean W-B *E* Difference	Mean W-B *E* Difference		Mean W-B *S* Difference[a]	Between *E*s Within Groups[b]
		White *S*s	Black *S*s		
K	.062	.062	.062	1.19	.252
1	−.199**	−.219**	−.180	1.37	.466
2	−.166**	−.192*	−.140	1.31	.523
3	−.011	.089	−.112	1.33	.369
4	.023	.295**	−.249*	1.63	.392
5	.080	.105	.056	1.75	.696
6	.065	.159	−.029	1.73	.517
Unweighted $\bar{X}$	−.021	.043	−.085	1.47	.459
Weighted $\bar{X}$	−.026	.047	−.088	1.47	.460

[a]All differences significant beyond .01.
[b]Not tested for significance.
*Significant at $p < .05$.
**Significant at $p < .01$.

TABLE 3
Lorge-Thorndike Verbal IQ: Mean Differences in Sigma Units

Grade	Mean W-B *E* Difference	Mean W-B *E* Difference		Mean W-B *S* Difference[a]	Between *E*s Within Groups[b]
		White *S*s	Black *S*s		
4	−.003	−.013	.007	1.59	.680
5	.404**	.371**	.437**	1.60	.415
6	.296**	.422**	.170	1.95	.710
Unweighted $\bar{X}$	.232**	.260**	.205**	1.71	.602
Weighted $\bar{X}$	.233**	.263**	.209**	1.71	.602

[a]All differences significant beyond .01.
[b]Not tested for significance.
**Significant at $p < .01$.

The overall mean difference between white and black *E*s (shown in the last two rows of the first column) is very small and nonsignificant even for these very large samples. The unweighted mean ($\bar{X}$) here is the simple arithmetic average of the means of every grade; the weighted mean is the average of the means of every grade, each weighted by the total number of *S*s on which the mean is based. Since the *N*s are usually similar from one grade to another, the weighted and unweighted means do not differ appreciably.

Columns 2 and 3 show the mean differences between white and black *E*s within each racial group of *S*s. Again, these differences are very small, and overall they are nonsignificant for the Nonverbal test. On the Verbal IQ (Table 3), both white and black *S*s perform significantly better with white than with Black *E*s. Any attempt to explain why the Nonverbal test shows nonsignificant *E* effects in grades five and six while the Verbal test shows significant effects, given the available information, would have to be sheer speculation. The Verbal and Nonverbal tests do not differ in the amount of *E* involvement in giving instructions, etc.

Column 4 shows the race difference between *S*s, against which one can compare the magnitudes of the differences shown in the other columns. The magnitudes of the racial group mean differences on the Lorge-Thorndike Verbal and Nonverbal IQ overall are some ten to 30 times greater than the magnitude of effects attributable to race of *E* × race of *S*s interaction.

The last column of Tables 2 and 3 shows the variation among *E*s within groups, that is, variation among *E*s not attributable to race of *E* or race of *S*s or the interaction of these variables. This variation among *E*s is expressed as the standard deviation of the means of *E*s within groups divided by the standard deviation of *S*s within groups. It should be noted, however, that some appreciable part of the variation among *E*s reflects differences between schools and classrooms, which inevitably results from the random assignment of a relatively small number of *E*s to a diversity of schools and classes. The interschool and interclass variations do not have a chance to "average out" over *E*s under the conditions of the present study. The between *E* variation allows no meaningful test of statistical significance but is presented here merely as a basis for comparing and evaluating the magnitudes of the other differences.

The above general comments serve as well for Tables 4 through 8.

Figure Copying Test

Results for the Figure Copying Test are shown in Table 4. The race of *E* effects can be seen to be quite small and unsystematic, though they are significant for white *S*s (in grades three and four) who do better with white *E*s. But the largest race of *E* effect (grade three) is less than one-fifth the magnitude of the mean racial group difference.

Speed and Persistence Test (Making Xs).

It is interesting that this test, which was devised to reflect *S*s' attitude and effort in a test situation and to be sensitive to motivating instructions, does in fact show far larger *E* effects than any of the other tests used in this study; differences amounting to half a standard deviation or more. It also shows by far the smallest overall racial difference between *S*s of any of the tests. The consistently significant race of *E* effects uniformly favor the white *E*s. The neutral and motivating instructions do not appear to produce

TABLE 4
Figure Copying Test: Mean Differences in Sigma Units

Grade	Mean W-B *E* Difference	Mean W-B *E* Difference		Mean W-B *S* Difference[a]	Between *E*s Within Groups[b]
		White *S*s	Black *S*s		
K	.002	−.015	.019	1.00	.317
1	−.076	−.009	−.144	.95	.343
2	−.048	−.079	−.017	.85	.420
3	.270**	.354**	.185	.87	.539
4	.078	.237**	−.081	.99	.521
Unweighted $\overline{X}$	.045	.098**	−.008	.93	.428
Weighted $\overline{X}$	.037	.085*	−.015	.93	.421

[a]All differences significant beyond .01.
[b]Not tested for significance.
*Significant at $p < .05$.
**Significant at $p < .01$.

TABLE 5
Speed and Persistence Test—First Try (Neutral Instructions): Mean Differences in Sigma Units

Grade	Mean W-B *E* Difference	Mean W-B *E* Difference		Mean W-B *S* Difference	Between *E*s Within Groups[a]
		White *S*s	Black *S*s		
1	.178**	.287**	.069	.56**	.812
2	.508**	.370**	.693**	−.09	.764
3	.542**	.588**	.496**	−.20**	.854
4	1.147**	1.185**	1.109**	−.44**	1.043
5	.550**	.650**	.450**	.10	1.062
6	.500**	.638**	.362**	.17**	.982
Unweighted $\overline{X}$	.571**	.620**	.530**	.02	.919
Weighted $\overline{X}$	.562**	.614**	.526**	.03	.915

[a]Not tested for significance.
**Significant at $p < .01$.

TABLE 6
Speed and Persistence Test—Second Try (Motivating Instructions): Mean Differences in Sigma Units

Grade	Mean W-B *E* Difference	Mean W-B *E* Difference		Mean W-B *S* Difference	Between *E*s Within Groups[a]
		White *S*s	Black *S*s		
1	.265**	.460**	.070	.53**	.793
2	.617**	.641**	.594**	.07	.727
3	.685**	1.013**	.357**	−.18**	.818
4	1.019**	1.263**	.775**	−.55**	1.086
5	.387**	.621**	.125	−.03	1.093
6	.477**	.630**	.324**	−.03	1.045
Unweighted $\overline{X}$	.575**	.771**	.374**	−.03	.927
Weighted $\overline{X}$	.570**	.766**	.374**	−.02	.921

[a]Not tested for significance.
**Significant at $p < .01$.

any differences with respect to the race variables. However, performance under the motivating instructions is significantly higher for all groups than under the neutral instructions. If the Making Xs test indeed reflects the kind of motivation that may enter into the attention and effort demanded by school learning, the present findings may have implications for possible race of teacher effects on scholastic achievement. This possibility would seem well worth investigation in its own right.

Listening-Attention and Memory for Numbers

These two tests were expected to show the smallest *E* effects, since their administration was wholly by means of a tape recording expressly intended to minimize variance due to *E*s. Administration of these tests involved the *E*s only as proctors and distributors of test forms. The expectation of small *E* effects is fully borne out by the results shown in Tables 7 and 8.

TABLE 7
Listening-Attention Test: Mean Differences in Sigma Units

Grade	Mean W-B *E* Difference	Mean W-B *E* Difference		Mean W-B *S* Difference[a]	Between *E*s Within Groups[b]
		White *S*s	Black *S*s		
2	−.055	.020	−.131	.23	.119
3	−.288**	−.039	.022	.36	.112
4	−.121	.036	−.279**	.32	.141
5	.059	.046	.071	.18	.461
6	.145*	.098	.192*	.19	.203
Unweighted $\overline{X}$	−.052	.032	−.025	.25	.207
Weighted $\overline{X}$	−.053	.031	−.030	.25	.203

[a]All differences significant beyond .01.
[b]Not tested for significance.
*Significant at $p < .05$.
**Significant at $p < .01$.

TABLE 8
Memory for Numbers Test: Mean Differences in Sigma Units

Grade	Mean W-B *E* Difference	Mean W-B *E* Difference		Mean W-B *S* Difference[a]	Between *E*s Within Groups[b]
		White *S*s	Black *S*s		
2	−.070	−.229**	−.125	.61	.114
3	.161*	.118	.204	.58	.201
4	.062	−.055	.180	.59	.215
5	.105	.063	.147	.67	.254
6	−.002	−.143	.139	.72	.236
Unweighted $\overline{X}$	.051	−.049	.109	.63	.204
Weighted $\overline{X}$	.048	−.055	.100	.63	.201

[a]All differences significant beyond .01.
[b]Not tested for significance.
**Significant at $p < .01$.

Conclusion

The magnitude of the race of *E* effect is better evaluated not so much in terms of its statistical significance but in relation to the magnitudes of other sources of variance in test scores and in relation to the size of differences in mental test scores that are of practical or theoretical consequence in any particular context. The present results on group-administered tests of cognitive abilities show unsystematic and, for all practical purposes, probably negligible effects of race of *E* on the mental test scores of the white and black school children. Moreover, the direction of the relatively slight race of *E* effects does not consistently favor *S*s of either race. The magnitudes of race of *E* effects are in all cases very small relative to the mean difference between the racial groups, except for the one noncognitive test, Making Xs, which is a measure of motivation or speed and persistence under the conditions of group testing. On this test, both white and black *S*s in all grades performed significantly and substantially (about 0.4 to 0.8 σ) better with white than with black *E*s. This shows that some types of performance are capable of systematically reflecting race of *E* effects and it tends to highlight the relative lack of such effects on the cognitive ability tests.

It should be noted that all *E*s were carefully instructed, trained, and supervised so as to insure as standardized and uniform testing procedures as possible. Only under such conditions can the race of *E* effects per se be properly investigated. The interest here is not in whether there could be race of *E* effects due to *E* differences in testing *procedures,* such as laxness in timing, individual variations and carelessness in giving instructions, and the like. To intentionally minimize variance due to nonstandard or sloppy testing procedures, the *E*s of both races in the present study were carefully selected and trained; they were equally competent testers.

How generalizable are the present results? This question, of course, raises the whole problem of judging the external validity of any empirical research. Strictly speaking, no result can be generalized beyond the specific populations which have been sampled, assuming strictly random sampling. To go beyond this (as nearly everyone does) is really not a question of statistical inference but a matter of scientific judgment. Since the present results are largely consistent with the failure of most other studies to demonstrate statistically significant race of examiner effects on cognitive ability tests, they probably have considerable generality. When a new finding contradicts a number of already established findings, the issues of level of significance and generalizability of the new findings, of course, become much more crucial. But other studies so far are not in conflict with the present conclusions, which merely add to the general consensus of the statistical evidence, though perhaps not to the consensus of popular opinion.

Since an entire school district was tested, the question, in a statistical sense, is not one of generalizing from a small sample to the larger population from which it was randomly drawn, but is rather a question of the kind of school population of which Berkeley may be typical and to which the findings in Berkeley may be generalizable. The factors most probably relevant to the present study are the progressive policies of the public schools in Berkeley and the climate of liberal attitudes and interracial interaction which encourages black participation in the schools. The school population in Berkeley, both whites and blacks, are above the statewide median for these groups in measures of scholastic achievement, and there is a larger mean difference between the groups. The district also has a higher average expenditure per pupil and employs a

larger percentage of minority teachers than the State of California as a whole. All testing for the present study was conducted a few months prior to the institution of busing as a means of achieving complete racial desegregation of the Berkeley schools. The majority of the 17 schools involved were predominantly either white or black because of residential patterns, but a few were already largely integrated prior to busing. If the geographical location and political climate of a school district affect the magnitude of race of *E* effects, this could be demonstrated only by conducting studies similar to the present one in widely differing communities. As yet, no one has done this.

Neither the preponderance of the evidence in the literature nor the results of the present study lends support to the popular notion that the race of the examiner is an important source of variance between whites and blacks on tests of mental ability.

REFERENCES

ABRAMSON, T. The influence of examiner race on first-grade and kindergarten subjects' Peabody Picture Vocabulary Test scores. *Journal of Educational Measurement,* 1969, *6,* 241–46.

BUROS, O. K. (Ed.) *The Fifth Mental Measurements Yearbook.* Highland park, N.J.: Gryphon Press, 1959, 478–484.

CALDWELL, M. B., & KNIGHT, D. The effect of Negro and white examiners on Negro intelligence test performance. *Journal of Negro Education,* 1970, *39,* 177–79.

CANADY, H. G. The effect of "rapport" on the IQ: A new approach to the problem of racial psychology. *Journal of Negro Education,* 1936, *5,* 209–219.

COSTELLO, J. Effects of pretesting and examiner characteristics on test performance of young disadvantaged children. *Proceedings of the 78th Annual Convention of the American Psychological Association,* 1970, *5,* 309–10.

DREGER, R. M., & MILLER, K. S. Comparative psychological studies of Negroes and white in the United States: 1959–1965. *Psychological Bulletin Monograph Supplement,* 1968, 70, No. 3, Part 2.

FORRESTER, B. J., & KLAUS, R. A. The effect of race of the examiner on intelligence test scores of Negro kindergarten children. *Peabody Papers in Human Development,* 1964, *2,* 1–7.

ILG, F. L., & AMES, L. B. *School readiness: Behavior tests used at the Gesell Institute.* New York: Harper & Row, 1964.

LA CROSSE, J. E. Examiner reliability on the Stanford-Binet Intelligence Scale (Form L-M) in a design employing white and Negro examiners and subjects. Unpublished master's thesis, University of North Carolina, 1964.

LIPSITZ, S. Effect of the race of the examiner on results of intelligence test performance of Negro and white children. Unpublished master's thesis, Long Island University, 1969.

MILLER, J. O., & PHILLIPS, J. A preliminary evaluation of the Head Start and other metropolitan Nashville kindergartens. Unpublished manuscript, Demonstration and Research Center for Early Education, George Peabody College for Teachers, Nashville, 1966.

PASAMANICK, B., & KNOBLOCH, H. Early language behavior in Negro children and the testing of intelligence. *Journal of Abnormal and Social Psychology,* 1955, *50,* 401–402.

PELOSI, J. W. A study of the effects of examiner race, sex, and style on test responses of Negro examinees. Unpublished doctoral dissertation, Syracuse University, 1968.

PETTIGREW, T. F. *A profile of the Negro American.* Princeton: D. Van Nostrand, 1964.

SATTLER, J. M. Statistical reanalysis of Canady's "The effect of 'rapport' on the IQ: A new approach to the problem of racial psychology." *Psychological Reports,* 1966, *19,* 1203–6.

SATTLER, J. M., & THEYE, F. Procedural, situational, and interpersonal variables in individual intelligence testing. *Psychological Bulletin,* 1967, *68,* 347–360.

SATTLER, J. M. Racial "experimenter effects" in experimentation, testing, interviewing, and psychotherapy. *Psychological Bulletin,* 1970, *73,* 137–160.

SATTLER, J. M. Racial experimenter effects. In K. S. Miller & R. M. Dreger (Eds.), *Comparative studies of blacks and whites in the United States.* New York: Seminar Press, 1973, 7–32.

SHUEY, AUDREY M. *The testing of Negro intelligence.* (2nd ed.) New York: Social Science Press, 1966.

AUTHOR

JENSEN, ARTHUR R. *Address:* Institute of Human Learning, University of California, Berkeley, CA 94720. *Title:* Professor of Educational Psychology. *Degrees:* B.A. University of California, Berkeley, M.A. San Diego State College, Ph.D. Teachers College, Columbia University. *Specialization:* Differential psychology, human learning.