HANKS, EDNA ELAINE. Examiner Race Effects on Standard vs. Culture-Fair IQ Performance and Correlations with Reading Achievement Test Performance. (1976) Directed by: Dr. Marilyn T. Erickson. Pp. 24.

The purpose of this study was to examine the effect of examiner race on, first, the IQ test performance of black male children with two types of tests, a standard intelligence test (Slosson) and a culture-fair intelligence test (CFIT) and, secondly, the correlation between these intelligence scores and reading achievement test scores.

The subjects were 40 black male first-graders attending public schools. The examiners were four black and four white female undergraduates who were trained in the administration of the intelligence and achievement tests for the purposes of this study.

Each subject was administered the Slosson Intelligence Test, the Culture-Fair Intelligence Test, and the Reading subtest of the Wide Range Achievement Test. One-half of each test (alternate items) was administered by a black examiner and the other half by a white examiner. Examiners were randomly assigned to subjects.

The data were analyzed using analysis of variance for a repeated measures all-within design and Pearson r correlations. No differences in IQ and achievement test scores due to examiner race were obtained. Subjects did, however, obtain significantly higher scores on the Culture-Fair Intelligence Test than on the Slosson Intelligence Test.

Correlation coefficients indicated that the highest correlations were obtained when both the intelligence tests and the Wide Range Achievement Test were administered by black examiners.

The findings of this study suggest that predictions about academic achievement for black male children may be better when tests are administered by black examiners. These results may have important implications for the future testing of black children.

EXAMINER RACE EFFECTS ON STANDARD VS. CULTURE-FAIR IQ PERFORMANCE AND CORRELATIONS WITH ACHIEVEMENT

by

Edna Elaine Hanks

A Thesis Submitted to
the Faculty of the Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Master of Arts

Greensboro
1976

Approved by

Marilyn T. Erickson
Thesis Adviser

APPROVAL PAGE

This thesis has been approved by the following committee of the Faculty of the Graduate School at the University of North Carolina at Greensboro.

Thesis Adviser Marilyn T. Erickson

Committee Members D. Evelyn LeSure

Jacquelyn D. Gaebelein

June 28, 1976

Date of Acceptance by Committee

ACKNOWLEDGMENTS

The author extends to Dr. Marilyn T. Erickson, her chairwoman, heartfelt thanks and gratitude for her guidance and aid in all phases of this project. To Dr. Jackie Gaebelein, and Dr. Evelyn LeSure, who gave suggestions and assistance, special thanks are also extended.

The author wishes to thank the administration, faculty, staff, and students of the Reidsville school system, without whose cooperation and interest this project would not have been possible.

The author appreciates the time and effort devoted to the project by her examiners: Lurlene Alston, Eva Carawan, Jamie Carlyle, Melodie Garner, Sue Gizinski, Lillian Johnson, Kathryn Johnson, and Evelyn Walser.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER I

INTRODUCTION

Testing has become one of the primary tools in the educational system. Many important decisions are made based upon test results, including placement, curriculum development, and suitability for careers. A child's entrance into the school system is begun with testing and throughout his school career various tests are called upon as sources of information. Thus, testing plays an important role in the child's life to the point of directing his future.

This important role necessitates our consideration of the factors that might influence the test results. Since the advent of the testing movement in the educational field, there has been awareness that such factors may greatly affect test performance. Examiner variables are one such group of factors.

The examiner in the testing situation is a potential critical factor and increasing research attention has been drawn to examiner variables such as age, sex, race, professional and socioeconomic status, appearance, and social warmth (Anastasi, 1961).

Examiner race is one factor that has been hypothesized to affect children's performance on standardized tests. This suggestion may be particularly important in view of the fact that most children (black and white) are tested by white

examiners who may not obtain optimum performance from all children. For example, behaviors such as fear and suspicion, verbal constriction, the assuming of a facade of stupidity in order to avoid appearing "uppity," and scoring low in order to avoid personal threat (Anastasi, 1958; Hilgard, 1957; Pettigrew, 1964; Riessman, 1962) have been reported by black examinees when tested by white examiners. It is questionable, then, that valid measures of intelligence are obtained for black examinees when the examiner is white.

Studies concerned with the race of examiner variable have been too few and inconclusive due to methodological problems (Sattler, 1966; Sattler, 1970; Sattler & Theye, 1967). Of the six recent studies done, four found the race of examiner variable to be significant, while two did not.

Abramson (1969) examined the effect of the race of examiner and the race of the subject on performance on the Peabody Picture Vocabulary Test (PPVT). The tests were administered by two white and two black examiners to 88 white and black first-graders and 113 white and black kindergarteners. The results indicated a small but significant interaction of examiner's race and subject's race for the first-graders but not for the kindergarteners. White children obtained higher scores with white examiners than black examiners, and black children obtained higher scores with white examiners. Abramson reported that all of the children were familiar with both the white and black examiners.

Caldwell and Knight (1970) used 15 male sixth grade black subjects and one white male and one black female examiner. The subjects were divided randomly into three equal groups and administered the Stanford-Binet, Forms L and M by a test-retest procedure. The black examiner administered Form L to Group A and Form M to Group B. The white examiner administered Form M to Group A, Form L to Group B and Forms L and M to Group C by the same test-retest procedure. They concluded that the race of examiner was not a critical variable based on their data which revealed no significant differences between groups. However, there were too few subjects and examiners used to warrant a firm conclusion. In addition, the authors did not include a comparison group in which both forms were administered by the black examiner. Also, race and sex were confounded in the study.

Pryzwansky, Nicholson, and Uhl (1974) suggested that it is also important to take into account the experiential background of the subjects. Thus, they compared 70 white and black females from rural and urban environments. Four white and four black examiners administered the Slosson Intelligence Test to a random sample of children. The results indicated that the race of examiner did not significantly influence the test performance of either rural or urban subjects. The author suggested that the sample size may have lowered the probability of obtaining a significant interaction between type of neighborhood, race of examiner, and race of subject.

Moore and Retish (1974) used a test-retest design to investigate the effect of examiner's race on the IQ scores of 28 male and 14 female low-income black preschoolers. The examiners, three white and three black females, administered an abbreviated form of the WPPSI (five subtests). Each subject was tested by a black examiner and a white examiner. Results indicated that the main effect of the examiner's race was significant for the Verbal, Performance, and Full Scale IQs. The children earned higher mean scores when tested by the examiner of similar ethnic origin. In addition, significant order of administration and order of administration x sex effects were revealed on the Verbal scale. The complete WPPSI should have been used although the abbreviated form does correlate highly with the complete form.

Solkoff (1972) used the WISC to examine the effect of examiner race on the performance of 112 black children and 112 white children, aged 8 to 11, with an equal number of males and females within each group. The examiners were four white and four black females. Each examiner tested 28 children, 14 black and 14 white, with an equal number of males and females within each group. The results indicated that scores on only one subtest, Information, were affected by the race of the examiner; both white and black children received higher scores when the examiner was black. The child's race and the examiner's race had significant effects on the Verbal, Performance, and Full Scale IQ scores. All

children received higher mean scores with black examiners on all three scales. He concluded that the findings indicated that white examiners do not depress IQ scores of black children, a conclusion totally unwarranted by his data.

In 1974, a portion of the study was replicated using 108 children, 54 black males and females and 54 white males and females. Two black and two white examiners administered the WISC. The examiner's race was a significant factor for the Comprehension, Digit Span, and Vocabulary subtests with the black examiners producing higher scores on the Comprehension and Digit Span subtests and the white examiners on the Vocabulary subtest. There was a significant race of child x race of examiner interaction on the Similarities and Object Assembly subtests. The white children achieved higher scores with a black examiner and the black children the higher scores with a white examiner. Again he concludes there is no evidence to support the idea that white examiners depress the performance of black children.

On the basis of the foregoing studies, it would be reasonably safe to conclude that the race of the examiner variable requires additional study. The critical issue here is not so much the IQ score <u>per se</u>, but the effect that examiner race has on the predictive validity of the test itself; that is, will the race of the examiner variable affect the ability of the test to predict future academic achievement? Further, we do not know under what conditions the

predictive validity would be enhanced or impeded. Since intelligence tests are generally used for predicting achievement, the question to ask is whether race of examiner impedes or enhances predictive validity.

The recent development of culture-fair tests also poses a problem for predictive validity. Thorndike (1971) states that the fairness of a test relates to the fair usage of the results. He defines fair use as follows: a common qualifying score may be used with two groups if the regression line based on one group does not systematically over- or underpredict criterion performance in the other. Cleary (1968) reiterates this point when she states that a test is biased for members of a subgroup of the population if the criterion score predicted from the common regression line is consistently too high or too low for members of the subgroup. Thus, a test or item of a test is biased for members of a particular group if it produces an uncommon discrepancy between the performance of that group and the performance of other groups.

Culture-fair tests were devised in an effort to eliminate this discrepancy. Research indicates that culture-fair tests correlate highly with standard tests of intelligence. The Culture Fair Intelligence Test, published by the Institute for Personality and Ability Testing, correlates .74 with the WAIS, .72 with the WISC, and .62 with the Stanford-Binet (Barton, 1973).

With the use of the culture-fair tests, again the question would be whether these tests are adequate predictors of

achievement? Further, as compared with standard intelligence tests, are the culture-fair tests better predictors when administered by black and/or white examiners?

There is insufficient evidence to determine whether culture-fair tests are, in fact, better measures of intelligence. Standard intelligence tests are not representative of the total population because of the limitations imposed by the standardization samples. The number of blacks included in the samples is not large enough to detect any real differences if such differences do, in fact, exist. A child who has been culturally deprived is not apt to respond with the acceptable response of the normative sample of which he is not representative. Even more important, will the specific cultural influences introduced by standard tests or the reduction of them introduced by culture-fair tests interact with the examiner race variable to enhance or impede predictability?

The first purpose of the present study was to examine the effect of examiner race on the IQ test performance of black children with two types of tests, a standard intelligence test and a culture-fair test. The second purpose was to examine the predictive validity of the two tests administered by black and white examiners. It was hypothesized that the children would perform better with examiners of the same race and that the predictive validity on both tests would be highest when the tests were administered by black examiners because black children would feel less threatened

by black examiners than with white examiners and because they are more accustomed to social reinforcement from blacks.

CHAPTER II

METHOD

Subjects

Forty first-grade black males, aged six to nine years, were selected randomly from eight classrooms in the Rockingham County school system in Reidsville, North Carolina. Reidsville schools were integrated in 1968. The black-white ratio for students is approximately 50-50 and 1 to 2 for teachers.

Examiners

Examiners were four white and four black female undergraduates at the University of North Carolina at Greensboro enrolled in Psychology 334, an independent study course. Female examiners were used in this study because examiners in school settings are typically female. Participation in training sessions, data collection, and once-a-week seminars were the requirements for the course. Examiners had no previous testing experience and were trained as a group by the author. The training was conducted over a three-week period with one and one-half hour sessions three days per week. Training consisted of a complete review of the test manuals, practice in reading the test items, and administration of the tests. The author observed the examiners administering the tests and provided immediate feedback. Examiners

were told that the purpose of the study was to compare two different types of tests and to determine whether testing procedures should be offered earlier in the curriculum for undergraduates. The examiners were given a complete explanation and the results of the study after data collection and analysis were completed.

Instruments

The standard test of intelligence used was the Slosson Intelligence Test for Children and Adults (Slosson, 1975). This test was selected due to its ease of administration and scoring and because it lends itself to half-test administrations. This test has adapted items from the Stanford-Binet, Form L-M, and thus correlates highly with it (r = .90 for age 4 to .97 for ages 18 and up). The standardization sample consisted of white, black, and American Indian children and adults from urban and rural sections of New York State. Items cover a wide range of abilities and are labeled by year and month of age.

The culture-fair test used was Scale 1 of the Culture-Fair Intelligence Test (Cattell, 1950). Scale 1 is designed for ages four to eight years and defective, institutionalized adults. It was selected due to its attractiveness to children and its freedom from the necessity of special abilities. Administration time is 40 to 60 minutes for a full test. The test consists of eight subtests: Substitution, Classification, Mazes, Selecting Named Objects, Following Directions,

Wrong Pictures, Riddles, and Similarities. The standardization sample consisted of 400 American and British individuals, with no significant difference between the two groups at any age level. This test correlates .72 with the WISC, .74 with the WAIS, and .62 with the Stanford-Binet.

The reading section of the Wide Range Achievement Test (Jastak & Jastak, 1965) was employed as the achievement instrument. The WRAT was standardized on children and adults in a number of states. Reliability coefficients of the reading subtest range from .981 for age 5 to .987 for ages 20 and up. The reading subtest correlates .80 for 47 children with the New Sanford Reading Test and .71 for 75 children with the WISC.

Procedure

Each subject was administered the Slosson, CFIT, and WRAT. Half (alternate items) of each test was administered by a black examiner and the other half by a white examiner. Examiners were randomly assigned to subjects and tested an equal number of subjects. Order of administrations for the type of test, race of examiner, and odd or even portion of the tests were counterbalanced. The intelligence tests were administered over a four-week period with only one type of test being administered within a week for a particular child. The achievement test was administered over a two-week period using the same one-half test procedure and randomization of examiners and odd or even portion of the test.

This procedure provided each subject with four IQ scores and two achievement scores. Achievement scores were converted to standard scores for comparative purposes. Intelligence quotients and standard scores were computed from the raw data by doubling the scores.

The author observed the individual administrations of the tests. Social contingencies were held constant by having examiners reinforce alternate responses, regardless of a right or wrong response, from a prescribed list of social responses (e.g. That's good, You're doing fine). Since there are data showing differential effects of reinforcement on IQ performance, this procedure was designed to insure that there were no differences between examiners due to this variable. Test manuals do not provide precise instructions with regard to social reinforcement. Scoring of the individual tests was performed by the author.

CHAPTER III

RESULTS

Table 1 presents the means and standard deviations for the three tests administered by black and white examiners. A 2 x 2 (race of examiner x type of IQ test) repeated measures all-within design using analysis of variance was employed. Table 2 shows the summary of the statistical analysis. The analysis of variance revealed a significant type of test effect ($p < .01$); subjects performed better on the CFIT than on the Slosson. No differences due to race of examiner were obtained. In addition, there was no race of examiner x type of test interaction.

An analysis of variance was employed for race of examiner for the achievement scores. Table 3 presents the data for this analysis. There was no difference on the achievement scores for the two groups of examiners.

Table 4 presents the correlation coefficients for the intelligence and achievement tests administered by pairs of examiners differing in race. All eight correlations were statistically significant. The highest correlation for the Slosson and the WRAT test scores was obtained when both tests were administered by black examiners. For the CFIT and WRAT test scores, the black-black examiner pairs also gave the highest correlation. While there were no significant differences between any of the correlations, a comparison

TABLE 1

Means and Standard Deviations of
Slosson IQ, CFIT IQ, and WRAT Achievement Scores
(N = 40)

Race of Examiner	Type of Test		
	Slosson	CFIT	WRAT
Black	M = 93.8 SD = 14.1	M = 106.8 SD = 25.3	M = 95.7 SD = 10.6
White	M = 96.0 SD = 17.0	M = 105.1 SD = 23.2	M = 97.1 SD = 10.1

TABLE 2

Analysis of Variance for Effects of
Examiner Race and Type of Tests on the Slosson and CFIT

Source of Variance	df	MS	F
Race of Examiner (A)	1	1.8062	.0113
Type of Test (B)	1	4917.3050	17.2094**
A x B	1	150.1523	1.0537
Subjects (S)	39	1078.1850	
A x S	39	159.8739	
B x S	39	285.7329	
A x B x S	39	142.5021	

** $p < .01$

TABLE 3

Analysis of Variance for Effects of Examiner Race on the WRAT

Source of Variance	df	MS	F
Race of Examiner (A)	1	37.8125	3.850
Subjects (S)	39	204.2682	
A x S	39	9.8111	

TABLE 4

Correlation Coefficients of
Intelligence and Achievement Scores

Examiner Pairs	Tests	
	Slosson and WRAT	CFIT and WRAT
Black-Black	.74**	.60**
Black-White	.66**	.56**
White-Black	.52**	.49*
White-White	.53**	.56**

**$p < .001$

*$p < .002$

of the correlations for black-black examiner and white-white examiner pairs revealed that the former accounted for 55% of the variance and the latter 28% of the variance when the Slosson and WRAT were administered. Differences of this magnitude were not found for administrations of the CFIT and WRAT.

CHAPTER IV

DISCUSSION

The present study did not obtain a race of examiner effect and is therefore compatible with the studies of Caldwell and Knight (1970) and Pryzwansky, Nicholson, and Uhl (1974). However, the results are not compatible with studies that did find a race of examiner effect (Abramson, 1969; Moore & Retish, 1974; Solkoff, 1972). Perhaps other conditions (e.g., female Ss, different age levels) would have shown an effect. This lack of a race of examiner effect may be accounted for by the fact that all examiners were drawn from the same population of undergraduates and received exactly the same training and supervision in test administration. In addition, the examiners had had no previous testing experience, and finally, social reinforcement during administrations was carefully controlled and presented in the same way by all examiners. The fact that all examiners received equal training, reducing the variance among them, may be highly important in the future testing of black children.

Societal changes may also have been involved in the findings. There have been substantial changes in school environments beginning with the Supreme Court decisions on desegregation. Black and white public school children not only attend school together but observe adults of both races

as teachers, principals, and other staff members. Furthermore, the black children in this study, being first graders, may not yet have had significant experiences with racial prejudice. That is, their relatively short experience in an integrated school environment had not sensitized them to racial differences.

Nevertheless, it is reasonable to assume that these children had had far more total experience with black adults as social reinforcers than white adults. Perhaps this experience served to reduce some of the error variance on the test administrations, thereby maximizing the validity of tests administered by black examiners. When the culture-fair test was used as the IQ measure, the race of examiner did not influence the validity of the CFIT. One possible reason for this effect is that the CFIT places less emphasis on verbal items.

The higher mean scores obtained on the CFIT neither increased the validity nor provided scores which were comparable (in terms of standard scores) with the WRAT. Not only was the predictive validity of the Slosson higher, but scores on the Slosson and the WRAT were more compatible.

The findings of this study suggest that, for black children, the Slosson is a more valid predictor of achievement when the tests are administered by black examiners. Future research should evaluate the race of examiner variable with the traditional tests of intellectual functioning,

namely the WISC and the Stanford-Binet. A recent study has suggested that the WISC may not be a valid predictor of achievement for black children (Goldman & Hartig, 1976). The major conclusion was that the WISC is a valid predictor for white children but not for black and Mexican-American children. However, their study did not include a standardized achievement test as a validity criterion, but rather relied on composites of academic grades and social maturity. With correlations of .25, .14, and .12, for white, black, and Mexican-American children respectively, one questions the strong conclusions drawn by the authors for any of the three groups.

Data on predictive validity should also be collected from older children to determine whether the present findings may be generalized to older children. High-risk children, who are "tracked" into educable mentally retarded classes, should also be the subject of this type of research utilizing "normals" as a comparative sample.

One problem in this area of research is that predictive validity may be interpreted to mean that IQ and school achievement tests both demonstrate the inability of the minority-group child to learn (Barnes, 1972). Such instruments do not measure the child's abilities to learn, but rather what has been learned in the context of the child's environment. Intervention studies with preschool culturally disadvantaged children have clearly demonstrated that IQ

scores and a variety of cognitive behaviors may be greatly improved (Heber & Garber, 1975).

This study adds relevant information about variables which affect black children's performance on tests. From the initial controversy of black versus white intellectual differences to queries about assessment measures, to examiner-examinee relationships and finally, and most importantly, to predictive validity, the testing controversy has delved into many important areas. However, increased concern should be most prevalent when one considers how the results are used in decision-making and the implications that these decisions have upon the child's future. Test results may influence adults' perceptions and expectations for children and, thus indirectly, influence how the child feels about himself. These expectations could possibly result in what is termed the "self-fulfilling prophecy," that is, the person tends to obtain what he predicts by "arranging" the situation so that his prediction is validated (Rosenthal & Jacobson, 1968). Because such important effects are associated with tests, it is critical to make certain that the test instruments and their usage are sound.

BIBLIOGRAPHY

Abramson, T. The influence of examiner race on first-grade and kindergarten subjects' PPVT scores. Journal of Educational Measurement, 1969, 6(4), 241-246.

Anastasi, A. Differential psychology. New York: MacMillan, 1958.

Anastasi, A. Psychological testing. New York: MacMillan, 1961.

Barnes, E. J. "Cultural retardation or shortcomings of assessment techniques?" In R. L. Jones (Ed.), Black psychology. New York: Harper & Row, Publishers, 1972.

Barton, K. Recent data on the culture-fair scales. Illinois: Institute for Personality and Ability Testing, 1973.

Caldwell, M. B., & Knight, D. The effect of Negro and white examiners on Negro intelligence test performance. Journal of Negro Education, 1970, 39, 177-179.

Cattell, R. B. Culture fair intelligence test: Scale 1. Illinois: Institute for Personality and Ability Testing, 1950.

Cleary, T. A. Test bias: Prediction of grades of Negro and white students in integrated colleges. Journal of Educational Measurement, 1968, 5(2), 115-124.

Edwards, A. L. Statistical methods. New York: Holt, Rinehart and Winston, Inc., 1973.

Goldman, R. D., & Hartig, L. K. The WISC may not be a valid predictor of school performance for primary-grade minority children. American Journal of Mental Deficiency, 1976, 80(6), 583-587.

Heber, R., & Garber, H. The Milwaukee project: A study on the use of family intervention to prevent cultural-familial mental retardation. In Friedlander, B. Z., Sterritt, G. M., & Kirk, G. E. (Eds.), Exceptional infant. New York: Brunner/Mazel, Volume 3, 1975.

Hilgard, E. R. Introduction to psychology. New York: Harcourt, Brace, 1957.

Jastak, J. F., & Jastak, S. R. The Wide Range Achievement Test. New York: Psychological Corporation, 1965.

Moore, C. L., & Retish, P. M. Effect of the examiner's race on black children's WPPSI IQ. Developmental Psychology, 1974, 10(5), 672-676.

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

Pryzwansky, W. B., Nicholson, C. L., & Uhl, N. P. The influence of examiner race on the cognitive functioning of urban and rural children of different races. Journal of School Psychology, 1974, 12(1), 2-7.

Riessman, F. The culturally deprived child. New York: Harper, 1962.

Rosenthal, R., & Jacobson, L. Pygmalion in the classroom: Teacher expectation and pupils' intellectual development. New York: Holt, Rinehart, and Winston, 1968.

Sattler, J. Comments on Cieutat's "Examiner differences with the Stanford-Binet IQ." Perceptual and Motor Skills, 1966, 22(2), 612-614.

Sattler, J. Racial experimenter effects in experimentation, testing, interviewing, and psychotherapy. Psychological Bulletin, 1970, 73(2), 137-160.

Sattler, J., & Theye, F. Procedural, situational, and interpersonal variables in individual intelligence testing. Psychological Bulletin, 1967, 68(5), 347-360.

Slosson, R. Slosson Intelligence Test for Children and Adults. New York: Psychological Corporation, 1975.

Solkoff, N. Race of experimenter as a variable in research with children. Developmental Psychology, 1972, 7, 70-75.

Solkoff, N. Race of examiner and performance on the WISC: A replication. Perceptual and Motor Skills, 1974, 39(3), 1063-1066.

Thorndike, R. L. Concepts of culture-fairness. Journal of Educational Measurement, 1971, 8(2), 63-70.