DOCUMENT RESUME

ED 048 383 TM 000 503

AUTHOR Findley, Warren G.; Bryan, Miriam M.

TITLE Ability Grouping: 1970 -- III. The Problems and Utilities Involved in the Use of Tests for Grouping Children with Limited Backgrounds, and Alternative Strategies to Such Grouping.

INSTITUTION Georgia Univ., Athens. Coll. of Education.
SPONS AGENCY Office of Education (DHEW), Washington, D.C.
PUB DATE Dec 70
NOTE 56p.
AVAILABLE FROM Dr. Morrill M. Hall, Director, Center for Educational Improvement, College of Education, University of Georgia, Athens, Georgia 30601. Identify the title and the part needed (Single copies)

EDRS PRICE EDRS Price MF-$0.65 HC Not Available from EDRS.
DESCRIPTORS *Ability Grouping, Culture Free Tests, *Disadvantaged Youth, Early Childhood Education, *Educational Strategies, Grouping Procedures, Heterogeneous Grouping, Individualized Instruction, Minority Groups, Standardized Tests, Student Grouping, Team Teaching, *Test Bias, *Testing, Test Interpretation, Test Reliability, Test Validity

ABSTRACT

Problems in the interpretation of standardized tests used to group children of limited backgrounds, cultural bias in tests, and the misuse of tests are considered. Reports on the use of specific tests with disadvantaged students are reviewed and some of the efforts being made to provide better interpretive data are discussed. Alternative strategies to homogeneous and heterogeneous ability grouping are suggested and described in some detail. The mutually compatible strategies include: individualized instruction, stratified heterogeneous grouping, student tutoring, team teaching, and early childhood education. An extensive bibliography and a list of test references are provided. See TM 000 501, 502, and 504 for other sections of this report. (PR)

Ability Grouping: 1970

III. The Problems and Utilities Involved in the Use Of Tests for Grouping Children with Limited Backgrounds, and Alternative Strategies to Such Grouping

Center for Educational Improvement
University of Georgia
Athens, Georgia 30601

FOREWORD

In December 1969, a task force was organized for the purpose of advising on the scope and organization of a series of reports regarding ability grouping in the public schools of the United States. Those involved in the planning included:

Warren G. Findley, Principal Investigator

Miriam M. Bryan	Edmund W. Gordon
Paul I. Clifford	Roger T. Lennon
John E. Dobbin	A. John Stauffer
Gordon Foster	Ralph W. Tyler

The Office of Education and the U. S. Department of Health, Education and Welfare were represented by Peter Briggs, Christopher Hagen, and Rosa D. Wiener.

Four documents were planned and have now been completed.

I. Common Practices in the Use of Tests for Grouping Students in Public Schools.

II. The Impact of Ability Grouping of School Achievement, Affective Development, Ethnic Separation, and Socio-economic Separation.

III. Problems and Utilities Involved in the Use of Tests for Grouping Children with Limited Backgrounds, and Alternative Strategies to Such Grouping.

IV. Conclusions and Recommendations

Mrs. Bryan prepared Document I, based on questionnaire responses from schoolmen and supplementary data from Miss Wiener. Dr. Clifford and Mr. Dominick Esposito prepared the basic content of Document II, which was then edited by Mrs. Bryan. Contributions to Document III were secured from Mrs. Bryan, Mr. Dobbin, Dr. Findley, Mrs. Blythe Mitchell, and Dr. Stauffer. The summary and conclusions were prepared by Dr. Findley.

The work presented herein was performed pursuant to a grant from the U. S. Office of Education, Bureau of Elementary and Secondary Education, Department of Health, Education, and Welfare. However, the opinions expressed herein do not necessarily reflect the position or policy of the U. S. Office of Education, and no official endorsement by the U. S. Office of Education should be inferred.

Additional copies of the four documents are available upon request. Write:

Dr. Morrill M. Hall, Director
Center for Educational Improvement
College of Education
University of Georgia
Athens, Georgia 30601

TABLE OF CONTENTS

INTRODUCTION

This document is presented in two parts: Part A is concerned with the problems and utilities involved in the use of tests for grouping children with limited backgrounds for purposes of instruction; Part B presents descriptions of alternative strategies. The first part was provided for in the outline originally set by the committee; the second part was added when the committee became impressed with the number of alternative strategies suggested in the literature as being effective and efficient. The strategies presented are merely representative of the variety of alternatives available. The reader may be able to add others.

A. THE PROBLEMS AND UTILITIES INVOLVED IN THE USE OF TESTS FOR GROUPING CHILDREN WITH LIMITED BACKGROUNDS

The search for useful information regarding the validity and reliability of standardized aptitude and achievement tests for use in grouping children with limited backgrounds for purposes of instruction has been an exhaustive but, unfortunately, not a very productive one. Not a single study, for example, among the more than two hundred located was found to involve all three aspects of the topic: test validity and reliability, culturally limited populations, and homogeneous grouping. It has been necessary, therefore, to attempt to go beyond the data presented and to make calculated inferences as to what might be expected to occur under certain combinations of circumstances.

Definition of Terms

The definition of a few terms is in order here if the intent of this document is to be clearly understood. These definitions may be read first or in conjunction with the discussion that follows. They are presented in a sequence of importance for understanding the material of this document. Wherever a term used in a definition is not understood, its definition is to be found later on.

1. In this document, concern will be for the validity not only of the tests themselves but also of their use for the whole population. Are the tests giving us the kind of information about students and about programs of instruction that we really want to know? In particular, do the tests provide comparable information about students with different backgrounds that can be useful in conducting the instructional program? Note particularly the definition of construct or pure validity given last.

The validity of a test refers to the extent to which a test does the job for which it is intended. Validity has different connotations for various kinds of tests and, accordingly, different kinds of validity are appropriate for them. For example, the validity of an achievement test is the extent to which the content of the test represents a balanced and adequate sampling of the outcomes(knowledge, skills, etc.) of the course or instructional program it is intended to cover (content, face, or curricular validity). The validity of an aptitude or readiness test is the extent to which it accurately indicates future learning success in the area for which it is used as a predictor (predictive validity). The validity of a personality test is the extent to which the test yields an accurate description of an individual's personality traits or personality organization as of that moment (status or concurrent validity).

The validity of a test or of a procedure for the use of a test for a particular purpose involves a combination of concurrent validity for indicating the present status of individuals in mastering a subject, predictive validity for indicating the probable later achievement of individuals in mastering that subject under specified instructional procedures, and freedom from correlation with extraneous variables on the part of the original or final measures of achievement. This total requirement may be called construct or pure validity. This concept of validity may be extended to other measures--self-concept ratings, personality measures, etc.--by substituting these terms for "test" in this definition.

2. The reliability of a test refers to the extent to which a test is consistent in measuring whatever it does measure: dependability, stability, relative freedom from errors of measurement. It is usually estimated by some form of reliability coefficient or by the standard error of measurement. The higher the reliability coefficient and the smaller the standard error of measurement, the more reliable is the test.

Reliability coefficients take their names from the method of determination. In this document we will be most frequently concerned with the alternative form coefficient, which is generally obtained by giving two parallel forms of a test (with equal content, means, and variances) to the same group of individuals on closely succeeding days and correlating the results; the split-half coefficient, which is obtained by correlating scores on one-half of a test with scores on the other half; the Kuder-Richardson coefficient, which is obtained from item statistics of a single administration of one form of a test; and the test-retest coefficient, which is obtained by administering the same test a second time after a short interval and correlating the two sets of scores. The alternate form estimate is generally preferred because it reflects the day-to-day variability implicit in ordinary use of tests.

3. The standard error of measurement is an estimate of the magnitude of the "error of measurement" in a score--the amount by which an obtained score differs from a hypothetical true score. It is the standard deviation of the differences between actual scores and theoretical true scores of the same individuals on a test. The standard error is an amount such that in about two-thirds of the cases the obtained score would not differ from the true score by more than one standard error.

4. A standard deviation is a measure of the variability or dispersion of a set of scores. The more the scores cluster around the mean, the smaller the standard deviation. It is the "root-mean-square deviation" originated by astronomers.

5. Correlation is the degree of agreement between two sets of data. In this document, the data will usually be scores on two tests for the same individuals, or scores on one test and marks given to the same individuals by a teacher. Less often they will be correlations between scores on other measures--interest inventories, personality scales, self-concept ratings--and test scores or marks.

Correlation is expressed in terms of a correlation coefficient, generally designated by the symbol r. This is an abstract number that can take on values between 0 and 1.00. The value of 1.00, almost never found, shows perfect agreement in the rank order of scores on one variable and scores on a second variable. The value 0, as that figure implies, shows absence of relationship between two sets of scores or random association between the sets. When the coefficient is preceded by a plus sign (+) or is presented without a sign preceding it, the correlation is said to be positive, with high scores on the first variable being most often associated with high scores on the second variable and low scores on the two variables also being associated with each other. When the coefficient is preceded by a minus sign (-), the correlation is said to be negative. This occurs less frequently, as one might expect, in that high scores on the first variable are most often associated with low scores on the second variable, and vice versa.

6. Multiple correlation is the degree of agreement between one variable, the criterion, and the best-weighted combination of a set of two or more other variables. An example would be the correlation between two test scores obtained at the beginning of a period of instruction--say, an achievement test score and an intelligence test score--and another test score at the end of instruction, generally an achievement test score in the same subject. A common example from outside the scope of this document would be the multiple correlation between high school average and entrance test scores used as predictors and grade point average at the end of the freshman year in college. Multiple correlation is expressed in terms of a coefficient of multiple correlation, designated by the symbol R to distinguish it from r, the symbol for simple correlation between two variables. This coefficient also takes on values between 0 and 1.00. When compared with the simple correlation between each of the predictor variables separately and the criterion, it shows the improvement in efficiency of prediction achieved by using the several variables in combination to predict the criterion. Multiple correlation R is always expressed without a sign because it can be used only to express the strength of a relationship.

7. A regression equation is an equation for predicting a criterion measure from the information provided by a single predictor or a set of two or more predictors. If a single predictor is used, we speak of simple regression or a simple regression equation; if two or more predictors are used, we speak of multiple regression, or a multiple regression equation. Correlation as described in definitions 5 and 6 preceding is the basis for determining the coefficients to be used in the equation.

Cultural Bias in Tests

The concept of cultural bias is receiving new attention. In the late 1940's and early 1950's much professional effort was devoted to analyzing tests with a view to producing "culture-free" or "culture-fair" tests (Machover, 1943; Turnbull, 1949; Davis, et al., 1951). Continuing efforts have been made by Cattell (1963) in his distinction between "crystallized" and "fluid" intelligence. Lorge (1952) pronounced a definitive evaluation of such efforts generally by pointing out that the major source of bias is to be found in society's "demands" and that tests must be related to those biases to define the cultural handicap of the disadvantaged in meeting the demands so that efforts may be directed toward correcting disadvantage and measuring progress in correcting it in individuals.

Two recent reviews, by Lambert (1964) and Anastasi (1964) merit mention as references here. Lambert summarizes information about a grest variety of measures of aptitude and achievement designed to be "culture-fair" and includes much obtained from direct correspondence or conversation with interested researchers. Anastasi clarifies the relations among a number of the measures and particularly the concept of culture-fairness as that varies with different groups studied and purposes served. For example, she points out that

> It is commonly assumed that nonverbal tests are more nearly culture-fair than are verbal tests. This assumption is obviously correct for persons who speak different languages. But for groups speaking a common language, whose cultures differ in other important respects, verbal tests may be less culturally loaded than tests of a predominantly spatial or perceptual nature.

Anastasi also points to factors that may normally be considered to limit the "culture fairness" of a test, but have validity in a particular situation. Thus

> . . .the same factor that lowered the test score would also handicap the individual in his educational and vocational progress and in many other activities of daily life. Similarly, slow work habits, emotional insecurity, low achievement drive, lack of interest in abstract problems, and many other culturally linked conditions affecting test scores are also likely to influence the relatively broad area of criterion behavior.

The reader should not be surprised, then, to find tests pronounced unbiased simply because they reflect the attributes that predict further achievement in school.

The view taken here separates society's demands into two chief parts: inescapable demands of living in an increasingly technological, urban, somewhat closed culture, and demands enforced by cultural distinctions of observable behavior largely associated with speech and historical knowledge. A current cigarette advertisement has capitalized on this by asking, "What do you want: good grammar or good taste?"

A common speech fault in English is use of the double negative, a "fault" generally reenforced for the disadvantaged child by the constant pressure of his home and neighborhood; yet in most modern foreign languages, the double negative is correct usage to achieve emphasis. And American students have to learn to correct their fault of forgetting to use the double negative!

Spelling is another mark of cultural bias. Among the readers of a publication like this, or of any publication intended for general currency, unfavorable notice would certainly be taken by many of faulty spelling if at all frequent. Yet it is doubtful that the meaning would have been unclear, as witness the fact that others will read by each error without noticing it. It may be noted that spelling enjoys the status of a schoool subject only in English-speaking countries because English is the only language not uniformly phonetic. Early emphasis on formal approaches to correct spelling can intimidate an otherwise competent child from exercising a free flow of writing for fear of misspelling. How much better a situation in which a child writes to inform distant parents that he has an "earake," enabling the family to swing into action immediately. "What do you want: good spelling or good medicine?"

The effect of frequent correction for the "stigmata" of poor speech and poor spelling is subject to review and curricular revision if it is agreed that early overemphasis on correctness produces academic and affective deficiencies. Certainly, there is a distinction now being pondered between society's cultural demands that all be able to read, calculate, communicate, and acquire a background of structured knowledge in order to participate effectively in society, and society's cultural biases which have been illustrated here from grammar and spelling, but which go much deeper.

Having made the above observations to put the matter of cultural demands in perspective, it is necessary to return to the earlier observations attributed to Lorge and Anastasi. The tests themselves as of any date must be judged in terms of their validity for predicting the currently accepted goals under current procedures of instruction.

The discussion that follows of Publishers' Test Information is limited to a sample of tests that are representative of the sorts frequently used in ability grouping at various grade levels from preschool to college. Considerable detail is given about a few tests widely used in elementary and secondary schools in grouping and in evaluating achievement. In addition, the most popular measure for use at the preschool level, a major college test and two new tests specially designed to meet the problems of testing minority children are discussed briefly. Thereafter the discussion proceeds in a subsequent section to relevant research studies of less specific emphasis.

Publishers' Test Information

The search for information about tests most widely used in school testing situations was initiated with a letter to each of seven major publishers of standardized tests asking for any data or other information they might have available about their own tests that would be pertinent to their use in ability grouping. Particular interest was expressed in predictive validity and/or reliability coefficients that the publishers themselves might have developed for groups differentiated by socioeconomic levels, or by race or ethnic background.

While only four of the seven publishers could provide useful data about tests on which they had done research, others reported research in progress, and all indicated that they were sensitive to the need for testing instruments free from cultural bias. Some reported the addition of members of minority groups to their professional staffs and provision for review of their test items by representative committees to detect instances of item bias.

Data supplied by test publishers are presented below. For some tests only reliability data are available; for others there are data regarding both reliability and predictive validity. With very few exceptions, these statistics show the tests to be unbiased with respect to any minority group, ethnic or socioeconomic; where such statistics favor one group over another, they appear to favor the minority rather than the majority group.

For the Preschool Inventory, formerly called the Caldwell Preschool Inventory, an instrument designed for use in the Head Start Program, Educational Testing Service reports deciles, summary statistics, and statistical characteristics for 317 children in eight kindergarten centers in North Carolina. This sample was divided into three groups by a consideration of each child's standing on two measures of socioeconomic status, the "Coleman" Index and an adaptation of the Ypsilanti Home Environment Scale, itself an adaptation of Wolfe's Environmental Process Scale. The two measures correlated .51 with each other. Scores for children at three socioeconomic (SES) levels increased from the low to the high group but the differences in mean score were not significant. KR_{20} reliability coefficients were .91, .89, .91, and .92 for low, middle, and high SES groups and the total group, respectively; for the total standardization sample the KR_{20} reliability coefficient was .91. Individual items which appeared to be unusually difficult or unusually easy for the low SES group were, more often than not, the same items that were unusually difficult or unusually easy for the total North Carolina group and for the standardization sample.

In the Directions Manual for the Clymer-Barrett Prereading Battery, published by Personnel Press, Inc., split-half reliability coefficients are presented for four groups of first-grade children selected because of their difference from the norming population or because they might present special testing problems resulting in unreliable work on the tests. These groups are described as follows:

Group A Kindergarten pupils tested in May; 120 children in 3 classes, one system. Mean total score 74.85.

-Richardson reliability coefficients, Formula 20.

Group B First grades in three-bi-lingual, rural schools in the Southwest; 63 pupils, mean total score 24.4

Group C First grade in a rural, white, low-ability school; 52 pupils, mean total score 20.0.

Group D First grade in a rural, Negro, low-ability school; 28 pupils, mean total score 24.2.

Group E Five first grades in two mixed-ethnic, deprived neighborhood schools in a very large city: 111 pupils, mean total score 25.6.

The reliability data for groups B, C, D, and E are presented below, together with those for the norms group. The data for Group A are omitted because they are for a group that is exceptional only in age (very young) rather in cultural background.

Table 1

Clymer-Barrett Prereading Battery

Reliability Coefficients for Special Groups and Norms Group

Test	Special Groups				Norms Group
	B	C	D	E	
Visual Discrimination	.96	.97	.94	.97	.94
Auditory Discrim.	.94	.98	.89	.94	.82
Visual-Motor	.91	.94	.95	.95	.89
Total (Short Form)	.94	.97	.93	.96	.92
Total (Full Form)	.97	.98	.96	.98	.95

The data indicate that even though the Clymer-Barrett Prereading Battery may be considerably more difficult for children in educationally atypical groups, it performs as well with them as it does with early first graders in the usual kinds of educational settings, so far as reliability is concerned.

By far the largest amount of data based on the use of tests with atypical groups has been published by Harcourt, Brace and Jovanovich, Inc. This is especially appropriate since their tests are used so widely in so many kinds of testing situations, especially those involving grouping.

For the Metropolitan Readiness Tests, the Manual of Directions provides split-half reliability data for seven different school systems at different socioeconomic levels with mean total scores ranging from 51 to 66. Since the

subtests are so short that it is recommended that relatively little significance be attached to the subtest scores of individual students, only the reliability coefficients for total score are shown.

Table 2

Metropolitan Readiness Tests

Split-Half Reliability Data for Form A in Seven School Systems

School System	Number of Students	Grade	Month of Testing	Mean Score	r^*_{11}
A	167	1	October	63.0	.91
B	173	1	October	57.9	.91
C	200	1	October	50.8	.94
D	88	Kdg.	May	66.4	.95
E	86	Kdg.	May	54.0	.93
F	59	Kdg.	May	53.4	.91
G	65	Kdg.	May	52.9	.90

*Indicates split-half reliability coefficient.

Table 3

Metropolitan Readiness Tests

Split-Half Reliability Data for End-of-Kindergarten Administration of Form B in Systems D, E, F, G

School System	Number of Students	Mean Score	r_{11}
D	82	66.5	.93
E	91	53.2	.94
F	55	55.8	.92
G	61	51.0	.93

Alternate form, or test-retest reliability data are also given for end-of-kindergarten children in systems D, E, F, G. For both Form A first-Form B second and Form B first-Form A second groups, total score reliabilities of .91 are reported.

With the observed reliability values for total score ranging from .90 to .95 and the measurement error of an individual score ranging from 3 to 5 points, as reported by the publisher, it would appear that total scores on the Metropolitan Readiness Tests may be used with considerable confidence for the purposes for which the tests are recommended.

The manual also provides predictive validity data for a variety of student groups and circumstances. The basic data include correlations between readiness scores and scores on the Stanford Achievement Test: Primary I (1964 Revision) the following May for 9497 students in the USOE First-Grade Reading Study of 1964-65 who participated in the standardization for the Readiness tests. Mitchell (1967) later used the scores of the same students to investigate the predictive validity of these tests and the Murphy-Durrell Reading Readiness Analysis by ethnic and socioeconomic differentiation. Certain of the Mitchell data, available upon request from the publisher, are summarized in Tables 4-6 on pages 10-12.

It is well to reiterate here the rationale of the statements above and below regarding bias in the tests. A test is adjudged to be biased only inasfar as it provides information that leads to faulty inferences. If a test gives dependable evidence of present status on a variable for members of a minority group, as measured by a high reliability coefficient, and if it also predicts subsequent achievement as well for minority groups as for the general population represented in the norms as measured by equally high correlation with achievement scores, the test is unbiased in its use for these purposes. The test may yield lower scores for minority group students, reflecting a disadvantage for the group on that test that is matched by the disadvantage these students experience in meeting the standard demands of instruction. Thus, the bias is in past conditions or in the absence of effective adaptation of instruction, rather than in the tests.

The results shown in Table 4 do not support the hypothesis that the Metropolitan or the Murphy-Durrell tests have lower predictive validity for minority group students than for white students. For the Metropolitan tests, of the 15 correlations shown, 12 favor minority groups; for the Murphy-Durrell tests, nine of the 15 correlations favor the minority groups. Nor is there any consistent pattern of advantage or disadvantage among the three minority groups.

Table 4

Correlations between Total Score on Metropolitan Readiness Tests and Murphy-Durrell Reading Readiness Analysis, Administered to First Graders in Early October, and Scores on Stanford Achievement Test Reading Subtests the Following May, for Various Ethnic Sub-Groups of the Total Group of 9497 Pupils Taking Both Tests

		Correlations of Metropolitan Readiness Tests with Stanford Achievement Test: Primary I, Form X					
Group	N	Word Reading	Paragraph Meaning	Vocabulary	Spelling	Word Study Skills	Standard Deviation of Metropolitan Readiness
White	7310	.58	.56	.59	.54	.59	15.8
Negro	518	.60	.55	.52	.56	.60	16.6
Mexican	139	.61	.56	.60	.57	.64	16.8
Oriental	37	.63	.51	.65	.60	.53	15.5
Ethnic origin unknown	1473	.68	.69	.69	.66	.71	19.3
Total Group	9497*	.63	.60	.63	.57	.64	17.5

		Correlations of Murphy-Durrell Analysis with Stanford Achievement Test: Primary I, Form X					
Group	N	Word Reading	Paragraph Meaning	Vocabulary	Spelling	Word Study Skills	Standard Deviation of Murphy-Durrell
White	7310	.60	.58	.52	.57	.59	26.7
Negro	518	.53	.56	.52	.58	.61	25.5
Mexican	139	.58	.55	.59	.58	.61	26.1
Oriental	37	.68	.58	.62	.62	.50	24.4
Ethnic origin unknown	1473	.69	.67	.63	.66	.69	29.9
Total Group	9497*	.64	.61	.57	.60	.64	28.4

	Standard Deviation of Stanford raw score				
Group	Word Reading	Paragraph Meaning	Vocabulary	Spelling	Word Study Skills
White	7.4	9.7	6.5	6.1	10.0
Negro	7.1	8.4	6.0	6.4	9.9
Mexican	7.1	8.6	5.8	6.3	9.8
Oriental	6.9	9.1	5.5	5.6	10.7
Ethnic origin unknown	8.7	10.7	7.3	6.9	11.8
Total Group	7.8	10.0	6.8	6.3	10.6

*The sum of the five Ns above is only 9477. The total group contained 15 Puerto Rican and 5 Indian-Eskimo children, for whom correlations were not computed.

Table 5

Metropolitan Readiness Tests

Predictive Validities of Total Scores by Adult Level of Education in the Child's Community

Median Adult Level of Schooling in Community	Stanford Achievement Test: Primary I, Form X					Standard Deviation of Metropolitan Readiness Score
	Word Reading	Paragraph Meaning	Vocabulary	Spelling	Word Study Skills	
9 years or less, N = 1411						
13 years or more	.57	.57	.59	.54	.57	14.4
9 years or less	.74	.70	.66	.64	.72	18.8
13 years or more, N = 1322						
Standard Deviation of Stanford raw scores						
13 years or more	7.4	9.9	6.4	6.0	9.5	
9 years or less	7.6	9.5	6.8	6.4	10.4	

Table 6

Metropolitan Readiness Tests

Predictive Validities of Total Scores by Median Annual Income of Community

	Above $8000, N = 1388			Below $4000, N = 1270		
Average Community Income	Stanford Achievement Tests: Primary I, Form X					Standard Deviation of Metropolitan Readiness Score
	Word Reading	Paragraph Meaning	Vocabulary	Spelling	Word Study Skills	
Above $8000	.647	.598	.601	.589	.626	14.5
Below $4000	.714	.700	.685	.630	.712	19.1
Standard Deviation of Stanford Raw Scores						
Above $8000	7.1	9.6	6.1	5.6	9.3	
Below $4000	7.7	9.6	6.7	6.5	10.6	

In terms of socioeconomic differentiation, the predictive validities of the Metropolitan Readiness Tests appear to be considerably higher for the scores of children in less privileged communities than for those in more privileged communities. In comparing the predictive validities in tables 5 and 6, however, it is important to consider the relative size of the standard deviations of the scores on the Readiness tests. The differences indicate greater variability for the readiness of children in the less privileged communities, and this would act to inflate the validities for these groups. Had the standard deviations for the two kinds of communities been more comparable, the differences in validities would have been less pronounced.

For the Otis-Lennon Mental Ability Test, also published by Harcourt, Brace and Jovanovich, Inc., split-half reliability data are provided for five socioeconomic levels of community. These are shown in Table 7 below.

Table 7

Otis-Lennon Mental Ability Test

Split-Half Reliability Coefficients for Socioeconomic Strata of the National Standardization Sample

		Otis-Lennon Level and Grade					Number of School Systems Within Stratum
		Primary I Grade 1	Elementary I Grade 3	Elementary II Grade 5	Intermediate Grade 8	Advanced Grade 11	
Socioeconomic Level*							
High	Median	.87	.90	.94	.94	.94	9
	Range	.79-.90	.87-.95	.90-.95	.92-.95	.94-.96	
Above Average	Median	.88	.94	.95	.94	.94	11
	Range	.85-.91	.90-.95	.94-.96	.92-.96	.93-.96	
Average	Median	.90	.92	.94	.95	.95	17
	Range	.87-.93	.87-.93	.83-.96	.93-.96	.92-.97	
Below Average	Median	.91	.92	.95	.95	.94	9
	Range	.88-.93	.89-.94	.94-.97	.92-.97	.93-.96	
Low	Median	.90	.92	.95	.96	.95	8
	Range	.89-.93	.90-.94	.93-.97	.93-.96	.92-.96	
Complete Standardization Sample		.90	.92	.95	.95	.95	

*Public school systems with less than 300 total enrollment were not included in this analysis.

In addition to the reliability data for different socioeconomic strata, the Technical Handbook accompanying the Otis-Lennon tests reports standard errors of measurement for successive score levels from IQ 50-70 to IQ 128-150. These range from 3.2 to 7.9 for single grades at single IQ ranges, from 4.4 to 6.6 for IQ level average, and average 4.9 for the total group.

Validity data for the Otis-Lennon test are reported for a large number of schools with mean IQs as high as 110 and as low as 94. Correlations between Otis-Lennon scores and scores on several widely used achievement test batteries and ability tests and with end-of-year course grades are given. School districts tested are identified as to SES level. Correlations between Otis-Lennon scores and scores on the achievement tests range from .50 to .80; correlations between Otis-Lennon scores and teacher grades are somewhat lower; and correlations between Otis-Lennon scores and scores on other ability tests are somewhat higher.

To aid in the interpretation of scores on the tests included in the College Entrance Examination Board Admissions Testing Program, the Board has published annually score report booklets for students, counselors, and admissions officers, and, periodically, much more comprehensive score reports. In addition, they have, through the years, commissioned a large number of research studies, and reports of many of these studies have found their way into professional journals. Two of these reports are particularly pertinent to the present discussion.

Studies conducted by Roberts (1962), Hills, Klock, and Lewis (1963), Boney (1966), and Stanley and Porter (1967) gave evidence that the Scholastic Aptitude Test (SAT) of the College Entrance Examination Board was as valid for predicting grades of students in predominantly black colleges as for predicting the college grades of white students (Kendrick and Thomas, 1970). The possible bias of the SAT in predicting college grades at integrated colleges was investigated by Cleary (1968) at the suggestion of the College Board.

Cleary and Hilton (1968) had earlier investigated possible bias in the Preliminary Scholastic Aptitude Test (PSAT) by studying the test items to see whether any items produced an uncommon discrepancy in scores for different racial and socioeconomic groups. On the basis of four separate studies of analysis of variance attributable to (1) "race," (2) SES, and (3) items, in the responses of 1410 twelfth-grade students who had taken PSAT in seven integrated high schools in three large metropolitan areas in 1961 (N = 636) or 1963 (N = 774), Cleary and Hilton concluded that while there were a few items producing an uncommon discrepancy between the performance of Negro and white students, the PSAT for practical purposes was not biased either for different ethnic groups or for groups at different socioeconomic levels. They based their conclusion on the absence of interaction* effects between item and "race" or item and SES.

*Interaction between two variables in an analysis of variance is a term to describe the tendency of individuals with particular combinations of status on the two variables to do much better or worse than would be indicated by their standing on the two variables separately. Here, if "race" or SES had given excessive disadvantage on particular items, the analysis of variance would have shown large interaction effects between item and "race" and/or item and SES.

The possible bias of the SAT is predicting college grades of black students at integrated colleges was investigated by Cleary (1968). She used the test as a whole as a predictor of college grade averages for both black and white students, hypothesizing that the test could be considered to be biased if too high or too low a criterion score was consistently predicted for members of the subgroup. Cleary concluded that there was no significant differences in prediction for black and white students from the two Eastern colleges represented in the study. At a third college in the Southwest, significant differences were found in the regression lines for black and white students, but it was a matter of overprediction of college grades for black students by the use of the white or common regression lines.

In a study parallel to Cleary's, involving 13 integrated colleges, Temp (in preparation) found that the use of a regression equation based on the majority or white student group resulted in the prediction of college grades for black students that were higher than those that they actually earned. According to Temp, colleges might consider the possibility of using separate regression lines for black students.

As this document is being written, a comprehensive technical report on research and development activities relating to the tests in the College Board Admissions Testing Program is in press (Angoff, ed.). In addition to an overview of administrative and technical problems of the program itself, the report describes construction practices involved in the Scholastic Aptitude Test and the achievement tests, discusses the statistical characteristics of the tests, the score scales, test validity, and the norms, and summarizes the results of several special studies having to do with the possible effect on test performance of coaching, test repetition, fatigue, anxiety, curriculum bias, and social and cultural factors. The Hilton and Cleary and the Cleary studies described above are among those reported.

A two-part Report of the Commission on Tests (College Entrance Examination Board, 1970) offers a variety of position papers, informed by research studies, on future directions for the College Board's program offerings. The commission of 21 members were drawn from persons variously concerned and qualified to deal with emerging issues in the use and interpretation of the tests in that program. The papers in this compilation, covering a broad range of purposes and services, bear in varying degree on the issues under discussion here. In particular, the opening article of Part II. Briefs, by John Carroll, endorsed by 19 of the 21 commission members, recommends revision of the widely used Scholastic Aptitude Test to accomplish better descriptive measurement of college applicants, especially the disadvantaged. Hope is expressed that psychometric techniques might be applied to the development of tests that will provide for separate report scores for (1) verbal knowledge (culturally influenced), (2) reasoning ability (largely verbal but less influenced by breadth and richness of cultural experience), (3) listening comprehension (a capability separately important and presumably less influenced by culture than reading), and (4) a de-emphasized section on quantitative reasoning (still hopefully allowing the culturally disadvantaged to show their potential as the present mathematics section does, relatively independent of verbal facility). The reader is directed to the original documents for the details which may be of particular interest and applicabllity in his own situation.

The American College Testing Program (ACTP), which seeks to serve the same function in college admissions, has its own intensive research studies in progress designed to identify item and/or test bias in its offerings. A major technical report, incorporating the findings of these studies, will likewise seek to map a course for the ACTP but is not scheduled for publication until late 1971 or early 1972.

Two new tests designed especially for use with the disadvantaged have recently been reported in the literature: A Reading Prognosis Test, published by the Institute of Developmental Studies, and the Orr-Graham Listening Test, also known as BoLt for Boys' Listening Test, published by the American Institutes of Research.

The Reading Prognosis Test is a 25-minute test, individually administered, measuring Language, Perceptual Discrimination, and Beginning Reading Skills. In a series of studies, the test was pretested and validated on balanced samples that included equal numbers of children from middle and lower socioeconomic groups and equal numbers of Negro and white children (Weiner and Feldmann, 1963). In an initial pilot study involving 40 children, the Reading Prognosis Test correlated .87 with the Gates Primary Reading Tests: Word Recognition of 1958. A second study involved 126 children, tested in October with the new test and in May with the Gates Primary Reading Tests: Sentence Reading and Paragraph Reading. In the October testing, retesting within three weeks of the initial testing yielded a reliability coefficient of .93 for the total group. At this time also the concurrent correlation with the Lorge-Thorndike Intelligence Tests for 138 children was .42 for the lower SES group and .21 for the middle SES group. The correlations of the Reading Prognosis total test score with the Paragraph Reading test ranged from .79 for the lower-class Negro female group to .89 for the middle-class white male group. The total group correlation was .81. The correlations of the Reading Prognosis total test score with the Sentence Reading test ranged from .61 for the middle-class Negro female group to .88 for the middle-class white female group. The authors concluded that the Reading Prognosis total test score, at the beginning of grade 1, is a good predictor of Gates scores for difference SES groups at the end of a year's instruction.

In a later validation study involving 300 Negro and white first graders in a large urban area and in a suburban community, correlations between the Reading Prognosis Test and the Gates Primary Reading Tests: Paragraph Reading and the Metropolitan Reading Test at the end of grade 1 ranged from .71 to .80, and correlations for separate ethnic and SES groups from .66 to .88 (Feldmann, 1965). Other and largely similar validation data are reported in the 1964-65 Research Memos of the Institute of Developmental Studies. Generally, the best prediction is shown to be for Negroes and for the lowest SES group.

The Orr-Graham Listening Test was developed between 1964 and 1968, with the financial support of the College Entrance Examination Board, to identify educational potential among disadvantaged eighth-grade Negro boys. The content of the test, an 86-item, 90-minute instrument administered orally, was designed to be of interest to boys of junior high school age. The stories in the test are based on such topics as spies, baseball players, cowboys and soldiers. The test was developed to elicit motivation through increased interest and to provide a test of aptitude which was not dependent upon reading proficiency.

All research, from that preceding the actual development of the test, through preliminary tryouts to the final administration, was carried on in junior high schools in the District of Columbia. About 99 percent of the boys included in the samples were Negroes. On the basis of a "final administration" of the test, Orr and Graham (1968) reported the test to be reliable, acceptable to the group for which it was intended, and uniquely different from the traditional aptitude and achievement tests. They obtained a split-half reliability coefficient of .85 and a Kuder-Richardson (20) reliability coefficient of .89. Correlations of the total test score with total scores on the School and College Ability Test (SCAT), STEP Listening, and STEP Reading were .60, .49, and .69 respectively. The results showed that about 81 per cent of the boys like the Listening Test and preferred it to a reading test covering the same content.

Carver (1969) reported on a replication of the Orr and Graham study with extension to other ethnic and income-level groups. In this study 615 eighth grade boys in the District of Columbia area, 314 Negroes (182 low-income, 132 middle-income) and 301 whites (110 low-income, 191 middle-income) were administered the Listening Test, SCAT (Level 2), and STEP Listening, and filled out questionnaires. Family incomes of $5000 divided the low- and middle-income groups.

An incidental reliability study of 142 low-income Negroes yielded an alternate form reliability of .78. For the low-income Negro group, correlations between the Listening Test and other test variables were highly similar to those in the earlier study; for all groups combined, the Listening Test correlated .69 with SCAT total score and .78 with STEP Listening, considerably higher than the correlations in the earlier study. The correlations between the Listening Test and STEP Listening ranged between .65 for the low-income Negroes and .79 for the middle-income Negroes. The low-income Negroes scored lowest on all tests, the middle-income whites scored highest on all tests, and the difference between these two groups was always greater than one standard deviation. The questionnaire responses showed that all four groups preferred the Listening Test to SCAT, but only the two Negro groups preferred it to STEP Listening.

Carver concludes that the reliability of the Orr-Graham Listening Test for low-income Negroes appears to be adequate and stable since there was little difference in the split-half correlations of the earlier study and the alternate forms correlations in his study. The concurrent validity is quite high, as indicated by the high correlation between the test and STEP Listening. The test also appears to be an adequate indicator of aptitude since the correlation with SCAT is high. He questions the high uniqueness of the test for identifying educational potential among the disadvantaged; to Carver the test is unique only in that it is preferred by Negroes. He finds no support for the hypothesis from the earlier test results that the effect of disadvantagement may be more associated with reading proficiency than with verbal proficiency in general. The large Negro-white differences are apparent in the Listening Test as well as in the reading and verbal measures.

In two other articles (1968, 1968-69) Carver further discusses the questionable uniqueness of the test and the failure of the test to lessen score differences between Negroes and Whites.

To summarize, systematic efforts are being made by test publishers and research agencies to review present test offerings and to introduce new emphases to meet the problem of assessing the capabilities of disadvantaged children. To date, the studies of old and new materials suggest possibilities but little accumulated capability for meeting the assessment problem directly.

The negative evidence that tests standardized on other populations tend to overpredict the subsequent performance of disadvantaged individuals, hence are not unfair to them, is cold comfort. The challenge is to mount a campaign of innovative teaching and evaluative research that will enhance learning by describing learning progress directly, rather than to settle for procedures that are fair only in the sense that they reflect "fairly" the current unmitigated disadvantages.

Now that the problem of assessing the potentiality and achievement of variously disadvantaged children is being faced, we must trust to continuing honest effort to separate the essential from the secondary objectives of public instruction to provide differential criteria of effectiveness of instructional adaptations. Thereby, it should be possible to help those operating from limited backgrounds to achieve increasingly greater mastery of essentials, including a self-respect that allows them to make a distinction between the essential and the ornamental outcomes of education.

Research Reports on the Use of Tests with the Disadvantaged

A second source of information, and a valuable one, was the Information Retrieval Center for the Disadvantaged at Teachers College, Columbia University. Useful studies found there were concerned with the testing of the culturally limited at all levels, from preschool to college students and adults; the testing of non-whites, including the Negro, the Mexican-American, and the American Indian; and the advantages and disadvantages of particular tests and particular types of tests for use with non-middle-class white groups.

Public libraries and university libraries gave access to the many periodicals in which articles were located through the Education Index, and to Dissertation Abstracts and Psychological Abstracts. The libraries of two test publishers proved a good source for unpublished studies. A visit to the Institute for Developmental Studies resulted in the location of other pertinent data, ERIC abstracts for reports related to disadvantaged and testing were examined.

Research relating to the effects of cultural background on test scores and the kinds of educational opportunities that have been afforded or denied the disadvantaged as a result of test performance has increased in volume and intensity as concern for the improvement and extension of opportunities generally for minority groups has become universal. But research of this kind is not new; for more than 60 years researchers have been exploring and reporting the complexities and problems of the use of tests with culturally different groups, even though for much of that time what they had to report may have been listened to by relatively few. While the great bulk of this research has been reviewed in preparation for the writing of this document, no attempt has been made to summarize the research that has been summarized elsewhere, except for those studies that have particular pertenence here. Instead, emphasis has been put

on those studies which have been done since 1960, most of them since 1965. Anyone interested in wider reading, particularly of the earlier studies, is referred to a half dozen of the most comprehensive surveys of the literature.

Lucas (1953) reviewed 253 pieces of literature relating to the effects of cultural background on scores on aptitude tests. Campbell (1964) included 46 references in his review of research done between 1932 and 1963 concerning the testing of culturally different groups. Pettigrew (1964) in the bibliography in his book on the Negro American listed among his 565 references almost 200 studies related to Negro-American intelligence. Shuey (1966) reviewed 382 studies in the latest edition of her volume bearing on racial differences in intelligence; while her conclusions relative to differences between Negroes and Whites, as determined by intelligence tests, have been the subject of considerable criticism, few would contest the statement that her coverage of the literature of the last 50 years is extensive. Dreger and Miller (1968) reported a comprehensive survey of psychological studies of Negroes and Whites done in the United States between 1959 and 1965. Flaugher (1970), in a recently completed review of research on testing practices, minority groups, and higher education, lists 65 references covering the years 1913 to 1970.

Studies of discrimination against minority groups in testing have usually dealt with the aspects of test content, the norms population, and the interpretation of results. What about the testing procedure itself? Do certain testing conditions systematically favor one cultural or racial group over another--examiner's race, test directions, pretest practice, speededness, test-wiseness? The next three studies were concerned with some of these conditions.

Pelosi (1969) made a study of the effects of examiner race, sex, and style on the test responses of adult Negro examinees. In his experiment, 96 Negro males were given six subtests of the Wechsler Adult Intelligence Scale (WAIS), the Purdue Pegboard, and the IPAT Culture Fair Intelligence Test, eight tests involving 12 scores, by examiners who included Negroes and whites, males and females, "warm" and "cold" personalities, with three examiners within each race-sex category. A separate analysis of variance was done for each of the 12 scores.

None of the examiner attributes or the interactions between them were significant on seven of the eight tests. The exception was the Culture Fair Test, group administered, for which "cold treatment by male Negro examiners resulted in substantially higher scores than those obtained by female Negro examiners." On all but one subtest of WAIS, the mean scores were higher with white examiners and for examinees treated coldly.

Pelosi writes: "Though differences were small and non-significant, the general direction contradicts the findings of previous research which suggested inadvertent negative bias due to white examiners." He suggests two weaknesses in the study, however: (1) The subjects were volunteers, enrollees in an anti-poverty work experience project, and were not as 'ego-involved" as would be the case in an actual testing situation. (2) The "warm" and "cold" examiners were not sufficiently different in the testing situations.

Abramson (1969) examined the effect of the race of both children and examiners on the child's performance on the Peabody Picture Vocabulary Test, an individually administered test. Two white and two Negro examiners administered the test to 88 and 113 white and Negro children in first grade and kindergarten, respectively, in an integrated urban school. The first graders had been in the school since their kindergarten year and the kindergartners had been in school for five months. The children had usually seen the examiner, a paraprofessional working in the school, at least once a day during the time they had been in school.

The investigator found a small but statistically significant interaction of the examiner's race and the child's race for first graders but not for kindergartners. He suggested that this difference might have been the result of the first graders having reached an age of racial awareness, but there were no data available regarding racial awareness.

A study reported by Dubin and Osburn (1969) was directed toward investigating whether two other conditions, aspects of the test procedure itself--extra preliminary practice and extra testing time--systematically favored white examinees over Negro examinees. Their sample included 235 Negro and 232 white students, representing both high and low socioeconomic levels, from two high schools in Galena Park, Texas. All students in the sample were quite familiar with standardized tests.

The Employee Aptitude Survey (four subtests) was used. Groups within each race in grades 9 and 10 were given the test with regular time limits; in grades 11 and 12 extra time was allowed. Some groups took only one form of the test; other groups took both forms, with the first testing considered as practice. An analysis of variance was done.

The order of mean scores was as follows:

By SES and Race	By Testing Conditions
High SES Whites	Power test with practice
Low SES Whites	Power test without practice
High SES Negroes	Speeded test with practice
Low SES Negroes	Speeded test without practice

Interesting findings of the analysis of variance were these:

1. Extra practice was no more advantageous to Negro than to white groups.
2. Both SES groups profited from extra practice to a comparable degree.
3. When Negro and white groups, matched by sex, grade level, and SES were compared, improvement in score from speeded to power tests was no larger for Negroes than for Whites.
4. High and low SES groups profited equally by the tripled time limits.
5. When both extra practice and extra testing time were given, again the improvement was not significantly related to either race or socioeconomic status.

The authors concluded that the results implied in a general sense that "testing procedure itself is not a major factor in discriminating between culturally advantaged and culturally disadvantaged students."

Goldstein et al. (1970) studied the effect of a specially designed enriched curriculum for 161 children on (1) average test performance over the two-year range from beginning pre-kindergarten to end of kindergarten, and on (2) stability coefficients over the same range for Stanford-Binet IQ, Peabody Picture Vocabulary Test, and the Columbia Mental Maturity Scale. Treating these three measures as measures of various aspects of cognitive development, they concluded that although mean gains on all three measures were reliable, the PPVT was not sensitive to effects of special instruction of these young disadvantaged children.

Lesser, Fifer, and Clark (1965) studied the influences of different social classes and cultures on patterns among mental abilities: verbal, number, reasoning, spatial. They tested 320 first-grade children, including middle- and lower-class Chinese, Jews, Negroes, and Puerto Ricans, in New York City and New Rochelle, New York, with the Hunter Aptitude Scales, designed for gifted four- and five-year-olds. Social class was based on the Hollingshead and Redlich index, using occupation, residence, and education of the head of the family as criteria. The scales were administered individually by well-trained psychometricians of the same ethnic group as the child.

Split-half reliabilities for the different ethnic groups (N = 80 for each group) ranged from a low .80 for Jewish children on Space to a high .96 for both Negroes and Puerto Ricans on Numbers. Split-half reliabilities by social class (N = 160 for each class) ranged from a low .80 for the middle class on Space to a high .96 for the lower class on Numbers. The middle-class children were slightly higher on Verbal but lower on Reasoning, Number, and Space. No tests for significance across ethnic or social-class differences were reported.

Means by ethnic group and social class are given below.

Table 8

Hunter Aptitude Scales

	Means for Ethnic Groups				Means for Social Classes	
	Chinese	Jews	Negroes	Puerto Ricans	Middle Class	Lower Class
Verbal	71.1	90.3	74.3	61.9	76.8	65.3
Reasoning	25.9	25.2	20.4	18.9	27.7	24.2
Number	27.8	28.5	18.4	19.1	29.8	25.6
Space	42.5	42.5	34.4	35.1	44.9	40.1

The greatest differences in standard deviation were in Verbal.

An analysis of variance was done, and interactions of social class, ethnic group, and sex reported. The major findings were that (1) differences in social class do produce significant differences in absolute level of each ability, but do not produce differences in the pattern of abilities; (2) differences in ethnic-group membership produce differences in both absolute level and pattern of abilities; (3) social class and ethnicity interact to affect the level of each ability, but do not interact to affect patterns. The authors concluded by proposing that "the

identification of relative intellectual strengths and weaknesses of members of different cultural groups become a basic and vital prerequisite to making enlightened decisions about education in urban areas."

Brazziel and Terrell (1962) conducted an experiment in the development of readiness in a culturally disadvantaged group of first-grade Negro children, most of them from sharecropper homes. Twenty-six of the children were assigned to an experimental group and the other 66 to three control groups. Parents of the children in the experimental group were involved in registration and in the development of readiness activities. The experimental group was given a six-week readiness program, which involved travelogues, 30 minutes of educational TV each day, and intensified activity to develop perception, vocabulary, and the will to follow directions. Weekly tests were given on some form of readiness.

At the end of six weeks, the Metropolitan Readiness Test was given to both experimental and control groups. The test results of the experimental group were greatly superior to those of the control group, the percentile rank for total score for the experimental group being 50 as opposed to 16, 14, and 13 for control groups A, B, and C respectively. The mean IQ of the experimental group in the spring of Grade 1 was 106.5, while second-grade Negro children in the country averaged 91.4 in the state testing program. Brazziel and Terrell attributed the success of the program to "an efficacious combination of direct teacher-parent partnership, excellent materials, test wisdom development, and energetic, uninhibited teaching. . . ."

Dowd (1969) studied sex and race differences in the effectiveness of various composite predictors of initial reading success and the relationship of children's self-perception to initial reading success. He tested 366 children from a large suburban district at the end of Kindergarten with the Metropolitan Readiness Tests (MRT), both the 1949 edition and the 1965 Revision, the Clark and Ozehosky U-Scale measuring self-concept, and the Van Alstyne Picture Vocabulary Test. At the end of Grade 1, he gave the Gates Primary Reading Tests: Word Recognition to 232 of the original 366 children still in school. For all groups (Negro, white- boys, girls) the best predictor was the MRT, except for the 1965 Revision for Negro boys; for them a combination of the Numbers and Copying subtests in the 1949 edition of the MRT provided the best prediction for the Gates tests. The U-Scale added significantly to the prediction on some instances; the Van Alstyne test did not.

Beidler (1969) worked with 276 students in Kindergarten through Grade 2 two schools in a disadvantaged neighborhood in Bethlehem, Pennsylvania, to determine the effects of the use of the Peabody Language Development Kits (PLDK) on the intelligence, reading, listening, and writing of disadvantaged children in the primary grades. The experimental groups had seven months of use of the kits in addition to the normal language arts program followed by the control groups.

The Lee-Clark Reading Readiness Test was administered to the Kindergarten in the spring, and the Otis-Lennon Mental Ability Test and the Cooperative Primary Tests in Reading and Listening to grades 1 and 2. A writing sample, scored for quantity and maturity, was obtained from grades 1 and 2.

At the Kindergarten level, there was a highly significant difference in favor of the control group, leading one to suspect that the experimental and control groups at that level may not have been initially comparable. For grades 1 and 2, no significant differences were found on intelligence, reading, or listening scores; in grade 2, however, the experimental group "wrote a significantly greater number of running words than did the control group."

Beidler described the implications thus: ". . .compared to conventional procedures, seven months of PLDK lessons do not significantly improve the intelligence, reading, listening, or writing of disadvantaged children in the primary grades."

In 1962 a study of socioeconomic status and school achievement was made by the California Elementary School Administrators Association. The School and College Ability Test (SCAT) and the Sequential Tests of Educational Progress (STEP) were given concurrently to 3008 sixth-grade students in 40 schools in three school districts. Grouping in terms of socioeconomic level (SES) was accomplished by use of the Hollingshead Two-Factor Index, based on parent occupation and education level. The two top groups, of five, were combined to make four SES levels.

Of pertinence here are the correlations between SCAT and STEP by SES levels. Was the prediction equally good at all levels?

The correlations between SCAT-Verbal, SCAT-Quantitative, and SCAT-Total and six STEP subtests by SES levels all followed the same general pattern. For all 18 sets of correlations, the lowest r's were for the highest SES level. For 11 sets of correlations, the highest r's were for the next to the lowest SES level. For none of the 18 sets of correlations, were the r's for the lowest SES level as low as those for the highest SES level. In other words, the prediction was generally better for the lower SES levels than for the higher SES levels.

The correlations between SCAT-Total and STEP by SES levels, from high to low, are given below.

Table 9
California Correlations between SCAT-Total and STEP by SES Level

		SCAT-Total						
STEP	N	Math	Science	Soc. Stud.	Rdg.	List.	Writing	S.D.* SCAT
SES A	524	.71	.62	.67	.64	.57	.61	10.7
B	566	.78	.72	.75	.72	.66	.70	11.3
C	524	.81	.78	.80	.76	.67	.74	9.0
D	553	.76	.74	.79	.77	.66	.69	7.6

* Standard deviation

Roberts et al. (1965) reported a longitudinal study of the performance of 69 Negro-American children on the Stanford-Binet Intelligence Scale, with special concern for the "causes or associated factors" of the observed differences. In this study different forms of the test were administered to the children at age 5 and age 10, with the second examiner having no knowledge of the earlier results. Data were gathered on parent occupation, family pattern, and socioeconomic level.

Over the five-year period, male mean IQ's fell from 96 to 88 and female mean IQ's from 94 to 84, with the decreases being statistically different in both cases. The respective standard deviations were 17.5 and 21.4 for the males, a large increase, and 13.2 and 15.4 for the females. The decline in IQ for boys seemed to be related to low socioeconomic status and unstable and unfavorable family patterns; the decline in IQ for girls was slightly in

reverse. The number of cases, however, was so small for the subgroups that little confidence can be placed in the statistics reported. The largest decreases were with children showing the greatest difficulty with verbal skills. Verbal Absurdities was an "outstanding failure." There was slightly less difficulty with Repeating Digits, and Making Change was relatively easy. None of the children tested at age 10 could pass the 10-year vocabulary test.

To obtain normative data on intelligence and achievement for a large homogeneous sample for which there were no previous data, Kennedy et al. (1963) administered the Stanford-Binet Intelligence Scale and the California Achievement Tests (CAT) to a well-selected sample of 1800 Negro students in grades 1 through 6 in five Southeastern states. They reported results by metropolitan, urban, and rural counties, age, sex, grade level, and socio-economic status.

For the entire sample the mean IQ was 80.7, with a standard deviation of 12.4. The mean IQ decreased with age, with type of community (from metropolitan to rural), and with socioeconomic level (from high to low); it remained relatively stable by grade. The order of the items by difficulty was quite similar to that of the norming population. The Negro students were relatively high on Rote Memory, Digits, Making Change, and Days of the Week, and low on Abstract Verbal, Vocabulary, Absurdities, and Comprehension.

On the CAT the mean grade equivalent on the total battery fell increasingly below the norm (from .2 in Grade 1 to 1.2 in Grade 5) and decreased with socio-economic level; there was, however, no difference in achievement by type of community. The correlation of the total battery with the Stanford-Binet mental age was .69, about the level usually found for total school groups.

Hughes and Lessler (1965) compared the Wechsler Intelligence Scale for Children (WISC) and Peabody Picture Vocabulary Test (PPVT) scores of 137 Negro and white rural school children of the lowest socioeconomic level in North Carolina. Ranging in age from 6 to 16, these children had been sent for testing because of suspected mental retardation. Could the shorter PPVT be substituted for the WISC, usually given?

Correlations between the two tests ranged from a low .21 for White Males for PPVT with WISC Performance to a high of .66 for Negro Males for PPVT with the Full WISC. Seven of the 12 correlations were .55 or higher. All but one of the r's was significant at the one per cent level and that one was significant at the five per cent level. Generally, the r's for Negro children were higher than for white children.

With the standard error of estimate* running from 7 to 14 points, the authors conclude that "the PPVT has a distinct advantage over group tests of intelligence for these rural children. . . and would perform an adequate screening function when used in the school or by personnel from the mental health clinic."

---

*The standard error of estimate is simply the standard deviation of the differences between scores of the same individuals on the criterion test and the predictor test, in this case expressed as IQ's. It is to be distinguished from the standard error of measurement, which accepts the test being studied as its own proper criterion and seeks to estimate departure of the value found on this test from the hypothetical true value that this test measures imperfectly because it cannot be infinitely long. See definition of the standard error of measurement on page 2.

Assign the children, particulary disadvantaged rural children, to EMR classes on the basis of a vocabulary test!

An investigation by Kneif and Stroud (1959) was planned, first, to provide data on the social class or culture bias in intellectual testing and, second, to ascertain interrelationships among certain relatively new intelligence tests and tests of scholastic achievement. The Lorge-Thorndike Intelligence Tests (L-T), Verbal and Nonverbal, the Davis-Eells Games, Raven's Progressive Matrices (RPM), and the Warner Index of Status Characteristics. All tests except the RPM were administered to a sample of 344 fourth-grade students in a Midwestern city, all the students present at the time in six of 18 elementary schools. One hundred sixty-four of these students who were in the fifth grade the following year were given the RPM.

All of the intelligence tests and composite scores on the Iowa Tests of Basic Skills (ITBS) correlated significantly with social status and, with the exception of the RPM, to approximately the same extent. The L-T Verbal scores gave the best prediction of ITBS scores, followed in order by L-T Nonverbal and the Davis-Eells Games. The L-T Verbal scores alone correlated with ITBS about as well as did the entire battery of tests when combined in multiple-correlation design. The RPM correlated to a smaller degree with ITBS than did any other intelligence test. The analysis gvae little justification for the use of L-T Nonverbal, the Davis-Eells Games, and RPM in conjunction with L-T Verbal for general prediction purposes. This is not to deny, however, their usefulness in individual diagnosis.

Davis (1969) followed 103 randomly selected students from Grade 3 through grades 5 and 6 to "measure improvement in test performance in disadvantaged inner-city poverty tracts" in Knoxville during a federally sponsored Communication Skills Project. The Metropolitan Achievement Tests of Reading, Word Discrimination, Language Usage,,and Spelling were administered in Grade 3 in 1967. Improvement was measured by relating to the 1965 results, 1966 and 1967 scores from California Achievement Tests in Reading Vocabulary, Reading Comprehension, Mechanics of English, and Spelling. Davis reports that "over the three test periods 48 comparisons for significance of differences. . .were run. Computed results indicated significant differences in thirty-two of the forty-eight comparisons."

Davis states in his thesis that "A basis for comparability of the MAT and CAT subtests was accepted when given correlation coefficients between areas of the two tests ranged from .77 to .95." It should be pointed out that correlation indicates only similarity in rank; it tells nothing of the grade equivalent scores, which could differ by months for students taking the two tests. There are also questions as to how standard scores and raw scores could be compared across the two tests (and levels) as the Grade 3 results on the MAT were compared with Grade 4 and Grade 5 results of CAT. Was "improvement" the gain from Grade 3 to later grades in the achievement areas considered? This comparison of results across different tests is very common even though not proper. There is evidence that MAT and CAT, particularly, are not comparable as to grade equivalent scores. CAT gives higher results and grade equivalent scores have a much smaller standard deviation.

The report appears to be attempted evaluation fo the effect of a federal projec:. How could this be measured by using gain over two years? There appears to be no relation of the gains to those of a group not in the study. What gains over the same period of time for the same schools had been made in previous years? What national norms give 1.0 as a normal yearly gain?

A study of Eagle and Harris (1969) examined the relationship between race and performance on two standardized reading tests, the reading tests of the Iowa Test of Basic Skills and the Metropolitan Achievement Tests. The tests were administered to 850 fourth-grade students and 850 sixth-grade students in all elementary schools of an urban district near New York City. Although white students earned higher scores than nonwhite students on both tests, the Metropolitan produced significantly greater differences between the races, at both grade levels, than did the Iowa. Ag Grade 4, the Metropolitan gave white students a superiority over nonwhite students of .72 compared to .58 for the Iowa. At Grade 6, however, the Metropolitan gave white students a superiority over nonwhite students of 1.13 years compared to .73 for the Iowa, a difference of about five months. Analysis of variance confirmed the statistical significance of these differences at both grade levels.

In brief, the Eagle-Harris findings imply that white elementary school children are "favored" by the Metropolitan whereas Negro children are "favored" by the Iowa when results are contrasted. Why is this so? Must one question the validity of one or the other of these highly respected tests? The authors suggest that in previous investigations involving comparisons among standardized achievements tests, little consideration has been given to the question of interaction effects between tests and sociocultural variables. Yet, failure to take into account significant interactions can mark important changes taking place in subgroup student performance and could provide the basis for erroneous or misleading evaluation of curriculum effectiveness.

The implications of findings like those of Eagle and Harris could be profound. With the knowledge that one test would be more reflective of gains for a particular subgroup than another, what administrator would not choose to use the test that demonstrates the kind of performance, maximal or minimal, that will best suit his practical purposes?

Santos (1967) studied the level and variability of achievement in educationally disadvantaged attendance centers in Iowa, and investigated item characteristics of the Iowa Tests of Basic Skills (ITBS) between educationally disadvantaged and total representative groups. In the Iowa 1966 testing program with ITBS, the educationally disadvantaged schools in all grades and all test areas were almost a year below the norm for representative schools. Difference in item difficulty between representative and disadvantaged schools was pronounced, and quite variable. The discrimination indices were equally satisfactory in the two groups. Santos suggests that research with experimental programs implies a need for reducing cultural bias, adapting content to needs and interests, and adjusting the difficulty of the test materials. "At the present time statements of behavioral objectives. . .are not specific enough to be of much help to authors of achievement tests in determining content, emphasis, and grade placement."

Buchanan (1969) studied the effect of cultural deprivation on the approach to test-taking as indicated by response style to multiple-choice questions. Buchanan asked whether his social background, deficient education, and experience of failure would lead the deprived student to reject the problem-solving approach when he is faced with questions to which he does not know the answers; that is, does he guess indiscriminately rather than attempt to eliminate the less plausible distractors in multiple-choice questions to arrive at an "educated" guess, as non-deprived students do?

Buchanan used three different tests at one grade level and one test at three different grade levels and analyzed (1) items on which non-deprived and deprived students experienced equal difficulty and (2) items with matched difficulty indices. For matched questions there was no difference between subcultural groups in the degree of selective guessing. Buchanan concluded that indiscriminate guessing is related to a real informational deficiency rather than to differences in motivation.

In a case study of the effects of educational deprivation on Southern rural Negro children, Green and Hoffman (1965) worked on the public schools of Prince Edward County, which were closed from 1959 to 1963. During these four years, most Negroes had no schooling (No Educ group); some had an average of one and one half years (Educ group).

After resumption of school operation, the Stanford-Binet Intelligence Scale and the Stanford Achievement Test-Partial Battery were given to 154 No Educ and 125 Educ. Extensive tables given by chronological age in the Green and Hoffman report show that the extended educational deprivation had a depressing effect upon achievement and intelligence at all ages. Language deficits on the Stanford-Partial were greater than in other areas. On the Stanford-Binet at the earlier ages (some children had never been to school), the differences between IQ's of children with No Educ and those with some Educ were as great as 30 points. In both the No Educ and the Educ groups, there was a negative relation between age and measured IQ.

Lo Monaco (1969) studied four groups of disadvantaged ninth-grade Negro boys to determine their response levels to both standard and oral-visual administrations of two vocationally relevant instruments. The boys were assigned to two experimental and two control groups equated for age, reading comprehension, and socioeconomic level.

Hypothesizing that reading deficits contaminate scores on standard versions of the instruments and that disadvantaged youth have better listening comprehension abilities than reading ability, Lo Monaco administered three measures--the Metropolitan Reading Test (MRT), the Kuder Preference Record-Vocational, and the Life-Planning Questionnaire-Modified (LPQ-M)--to all groups in the standard version and in a modified oral-visual version involving no reading. The two experimental groups took both the standard version and the oral-visual version in difference sequence; one control group took the standard version twice, and the other the oral-aural version twice.

Except for the Reading Test, oral-visual version scores were higher than the standard version scores on all measures; on the MRT, this was true for the low reading cases only. The oral-aural version provided more reliable [illegible]res of interests on the Kuder and of strivings on the LPQ-M than did the [illegible]ard version.

According to Lo Monaco, "the findings of this study indicate that reading deficits are important response variables. . . ." Instruments can be modified to "mediate these difficulties."

Alzobaie, Metfessel, and Michael (1968) administered the Lorge-Thorndike Intelligence Tests, Verbal and Non-Verbal, three of Guilford's tests of creativity, the Test of Academic Performance-Reading, and two scales from the Cattell Culture Fair Intelligence Test to 122 disadvantaged tenth-grade Negro students, in a district adjacent to Watts in Los Angeles. Grade point averages (GPA) and SES indices from the Warner Index of Social Class scale were also obtained for each student.

Intercorrelations among the predictors ranged from .23 to .82; the Guilford total score had correlations ranging from .40 to .56 with the other predictors. The Lorge-Thorndike and Reading tests showed small but significant correlations with SES; the Guilford and Cattell tests did not. Correlations with a convergent criterion measure* of academic success--GPA ranged from .29 and .32 for the Cattell scales to .56 for the Reading test; correlations with GPA for the three Guilford tests, essentially divergent tests, were .46, .39, and .31, with .48 for the composite.

The authors conclude:

> Despite their brevity, the three essentially non-verbal tests of divergent production as well as their composite score showed promise in the prediction of GPA. Thus, the three Guilford tests afford an alternative means for predicting traditionally evaluated academic performance of culturally disadvantaged children, many of whom have substantial disabilities in both receptive and expressive language function relative to expectations of a middle-class Anglo-American culture.

Harris and Lovinger (1968) investigated the commonly reported tendency of Negro IQ's to drop with increasing age in a longitudinal study involving 35 boys and 45 girls in a very disadvantaged area in the borough of Queens, New York City, in a school which had the lowest achievement and highest transiency rate of any junior high school in the borough. All 80 students had been given the same tests from the first grade on: Grade 1, Pintney-Cunningham Primary Test; Grade 3, Otis Quick-Scoring Mental Ability Test: Alpha Level; Grade 6, Otis Quick-Scoring Mental Ability Test: Beta Level; Grade 7, the Wechsler Intelligence Scales for Children (WISC); Grade 8, the Cattell Culture Fair Intelligence Test and the Pintner General Ability Test; Grade 9, WISC. There were 12 measures in all.

No decrease in IQ was found throughout successive grades for this group of disadvantaged Negro adolescents. Mean IQ at Grade 1 was 98, then 94, 88, 93, 96, 92, to 96 at Grade 9. On the WISC this group was not any more handicapped on verbal than on non-verbal tests. At Grade 7 the mean was 93.8 for Verbal and 93.7 for Performance; at Grade 9 the means were 96.1 and 97.0, respectively. The correlations between the tests given two years apart were .87 for Verbal, .85 for Performance, and .89 for Full Scale.

*The authors write: "Time limits of convergent tests favor the time-conscious middle-class culture."

The purpose of a study by Bradley (1967) was to investigate selected characteristics, academic performance, personal problems, and successes of Negro undergraduates in seven formerly all-white Tennessee colleges and universities. In addition to course grades, personal and social data were collected on 583 students over a two-year period of means of interviews and a student questionnaire.

One result is pertinent for reporting here. The multiple regression equation for best predictions of grade point average (GPA) includes these variables in this order: (1) high school GPA, (2) a confidence in ability factor, (3) the American College Testing Program (ACTP) social studies score, and (4) a morale factor. The multiple R predicting college grades was .6131, with a standard error of estimate of .5451 (one half the difference between two letter grades, ad C and B).

Interestingly, Bradley found that no ACT score other than that for social studies added any significant increase. In Bradley's words: "The ACT scores in English and math cannot be used as a basis for predicting the academic success of the Negro students in the same way that they are used to predict college success for privileged white students."

Boney (1966) studied 104 Negro boys and 118 Negro girls in Grade 12 in a Port Arthur, Texas, high school. The Cooperative School and College Ability Test (SCAT) had been given in Grade 8. Three subtests from the Differential Aptitude Tests were administered at the end of Grade 12, concurrent with the computation of the grade point average (GPA). A multiple correlation of .80 for boys and .82 for girls resulted when the predictors of junior high school grade point average, the Sequential Tests of Educational Progress (STEP) in Language and Social Studies, the California Test of Mental Maturity, and the three DAT subtests wree combined. Because 97 per cent of the parents were unskilled laborers, there was little discrimination in socioeconomic status (SES) and SES did not become part of the regression equation. Boney concluded that "Negro students are as predictable as other groups" and that "prediction could be made in junior high school."

Wilson (1969) reported a study undertaken by College Research Center in order to facilitate the efforts of a group of eight highly selective liberal arts colleges for women to evaluate the progress of black students enrolled at the time and to develop rationales for extending educational opportunity to members of disadvantaged minority groups. The study focused on (a) selected characteristics of black women who entered member colleges of the College Research Center in 1965, 1966, and 1967, and (b) the correlational validity of standard admissions criteria for predicting college grades.

Black students entering CRC--colleges during the study, themselves a select group, differed from their classmates in a varietyof educationally relevant ways--in socioeconomic background, career orientations, perceived purposes of college, educational plans, attitudes, and in level of performance on standard admissions variables (measures of academic aptitude, SAT Verbal and Mathematical), scores on College Board Achievement Tests, and secondary school standing. The findings of the study suggest that, despite such differences, forecasts of freshman-year academic performance are likely to be at least as accurate for black students as for their white classmates. There is, moreover, some evidence that predictions made on the basis of standard formulas may tend to overestimate the first-year performance of black students in the several colleges studied.

"It is commonly assumed that scholastic aptitude tests are based against culturally different or disadvantaged students. . .but it is important to know whether they have useful validities for predicting relative criteria for such students." So wrote Munday (1965), who studied the predictive value of the American College Testing Program (ACTP) for 1658 students in five 4-year Negro colleges in four different Southern states. Munday employed five separate criteria (college English, mathematics, science, social studies, and overall averages). He found that the multiple R's derived from optimally weighting four high school grades in each category was lower than the multiple R's derived from the optimal weighting of the four ACTP tests. The latter R's gave predictions of college grades that were as good for the Negro colleges as for all colleges using the ACT service.

Munday described his findings as being consistent with those from other studies, that is, that grades for socially disadvantaged students are generally as predictable as grades for other students using standardized measures of academic ability. In Munday's words: "If such tests are culture-bound, as seems likely, this feature does not appear to detract from their usefulness as predictors of academic success."

Mexican American Studies

In one of a series of studies investigating the possible bias of testing Spanish-speaking children in English, Davis and Personke (1968) gathered evidence concerning the effects of administering the Metropolitan Readiness Test (MRT) in English and Spanish to 88 Spanish-speaking children in their first school year in a South Texas city. Fifty-three of the children were enrolled in pre-first grade sections, or "readiness classes" designed for children deficient in the English language; 35 of the children were in regular first-grade sections. Early in the school year, the Spanish version of the MRT, with published test directions in English translated into South Texas colloquial Spanish, was administered to all of the children by the same individual, and the English version, according to school practices, by the classroom teachers. Contrasts of mean differences on subtest and total scores on the two modes of test administration yielded mostly non-significant differences. The children performed at a significantly higher level on the subtests on Word Meaning when the test was administered in Spanish; on the subtests on Alphabet and Numbers, however, significant differences favored the administration of the test in English. The findings did not show that administration of the MRT in English rather than Spanish resulted in any inadequate assessment of and substantial testing bias against Spanish-speaking children.

As a second phase of the study, Personke and Davis (1969) administered the Metropolitan Achievement Tests (MAT) in May to the first graders who had participated in the earlier testing with the MRT. The total score on the English administration of the MRT was a significantly better predictor of performance on the Word Knowledge subtest of the MAT than was the total score on the Spanish administration. For the other two subscores on the MAT, Word Discrimination and Reading, the English administration of the MAT yielded higher, but not significantly different, coefficients of correlation than the Spanish administration did. Of 12 comparisons made between the subtests of the MRT (English and Spanish versions) and the three scores on the MAT, six differences were statistically significant, and these differences divided themselves equally

between the English and Spanish administrations. The administration of the MRT in English rather than in the children's native Spanish apparently did not result in test bias for these children.

While the results of this research are interesting and impressive, one wonders how any other outcomes could have been anticipated. If children are being taught to read English, then their readiness to learn should be best assessed in terms of their ability to cope with the English language; and the greater that ability, the greater the amount of progress in reading achievement to be expected.

Karabinus and Hurt (1967) described the results of the revised Van Alstyne Vocabulary Test given to 535 six-year-old Mexican-American children attending poverty-qualifying schools in Tucson, Arizona. Spearman-Brown, Kuder-Richardson, and test-retest reliability coefficients for the scores of the Mexican-American children ranged from .76 (Kuder-Richardson) to .87 (test-retest), as compared with .71 (Spearman-Brown) for the general norming population. Concurrent validity coefficients with the Stanford-Benet Intelligence Scale, the Wechsler Intelligence Scale for Children, and the Metropolitan Readiness Tests, were above .60. While the Van Alstyne test was judged to be both reliable and valid for the measurement of mental ability of these Mexican-American children, the mean mental age for the two groups was so much lower than that of the general norming population (33.4 as opposed to 44 to 47) that a normalized frequency distribution of raw scores showing corresponding percentile ranks was developed for use with the Mexican-American children rather than the percentile ranks for IQ scores provided in the manual. It was suggested that the special norms might be useful when measuring other culturally disadvantaged children.

Morper (1967) studied the relationship between certain predictive variables and achievement measures for Spanish-American and Anglo ninth graders in Oklahoma. To 50 children of each ethnic group he administered the Wechsler Intelligence Scales for Children (WISC), the Lorge-Thorndike Intelligence Tests, and the School and College Ability Test (SCAT) as predictive measures. Achievement measures included teacher marks in English, mathematics, and science and the Metropolitan Achievement Tests.

For the Spanish-American group, neither the WISC nor the Lorge-Thorndike IQ's correlated at the 5 per cent level of significance with scores on the MAT; while for the Anglo group, all three predictor variables correlated satisfactorily with the MAT scores. With teacher marks as criterion variables, the correlations for all predictive variables were significant for both ethnic groups. The greatest differences between the Spanish-American and Anglo groups were observed when reading ability and comprehension were most involved in the obtaining of a measurement, the difference being in favor of the Anglo group.

Kimball (1968) studied parent and family influences on the academic achievement of Mexican-American students. His population included 1457 Grade 9 students from eight junior high schools, 899 Mexican Americans and 558 Anglos. Twenty-three variables were tested for association with (1) school marks, (2) achievement test scores, and (3) general ability. Parental educational aspirations for their child was significantly related to all achievement variables and was more strongly related to achievement than were personal identity, background, family structure, social status, and ethnic status. Just below parent influence in predictive ability were per cent of Anglos in the school, socioeconomic status, father's education, family

intactness, family birth in Mexico, grandparents' residence, and birthplace of child. Sex, age, birth order in family, and family size were of little consequence.

A comparison of Mexican-American and Anglo patterns of relationship between achievement and these independent variables were found by Kimball to indicate more overall differences than similarities.

Chandler and Plakos (1969) of the Mexican-American Education Project conducted an investigation to determine whether certain Mexican-American students belonged in Educable Mentally Retarded (EMR) classes or whether a language barrier prevented them from being assessed properly as to their native abilities to perform cognitive tasks. Their sample included 47 students of Mexican descent, with a problem in using the English language, in grades 3 through 8 in two school districts, an urban and a rural district, in different geographical areas.

The Spanish version of the Wechsler Intelligence Scale for Children (WISC) was administered and scores interpreted in terms of norms developed in Puerto Rico. (Because this version was in Puerto-Rican Spanish, some items had to be reworded and some changes made in the key.) The IQ's so obtained were compared with previous IQ's based on a test not identified. The mean IQ gain was 12.4, with 44 of the 47 students scoring higher on the Spanish WISC. The median IQ was 83, as compared with a median IQ of 70 on the test administered earlier. Only 9 of the 47 scores were below the cutoff IQ of 75 for EMR classes when the Spanish WISC was given.

Of interest to note here is an experiment conducted by Palomares and Johnson (1966) that demonstrated the crucial role played by the psychologist in the over-representation of Mexican-American children, or, for that matter the overrepresentation of children of any minority group, in EMR classes. Palomares and Johnson each tested and interviewed approximately 35 Mexican-American children, ages 7 to 14 years, who had been recommended for EMR class placement. After testing the children with the Wechsler Intelligence Scale for Children (WISC), the non-Spanish-speaking psychologist, Johnson, found 24 of his 33 students, or 73 per cent, eligible for EMR classes, while the Spanish-speaking psychologist, Palomares, recommended that only nine of his 35 students, or 26 per cent, be placed in EMR classes. Clearly examiners, as well as tests, can differ even when the students tested are similar and the test used, the same. There is little doubt but that a larger scale experiment would result in similar findings. Incidentally, both examiners averaged IQ estimates of 95 on the Goodenough-Harris Draw-a-Man and Draw-a-Woman Test for children on subsamples of 25 for whom the WISC total IQ's averaged 70 and 75, respectively.

Metfessel (1965) studied attitude and creativity factors related to achieving and nonachieving disadvantaged youth, largely Mexican-American. He found that Individual Tests of Creativity are considerably superior in predicting the academic behavior generally and of Mexican Americans particularly, than traditional measures of intelle t and scholastic aptitude. Correlations of the scores on these creativity tests with grade point averages were ranging from .39 to .49 at the time Metfessel reported. The Inventory of Self Appraisal and the Meaning of Words Inventory, two relatively independent tests of the achievement motive, were correlating between .36 and .44 with grade point average. Metfessel concluded that the results appeared to indicate that "the above three tests combine to produce a potent unified approach to forecast student achievements."

The eight Mexican-American studies briefly annotated in this section cover thinly the same general issues treated more fully for blacks and whites of low socioeconomic status in the preceding sections. The added feature is the foreign language component; ghetto children suffer language handicaps, but nothing quite as "wrong" as a wholly different language base. The Palomares-Johnson difference of interpretation of essentially the same low performance on individual tests is an echo of the Kariger (1962) finding reported in the previous document that personal judgment compounds the ethnic separation produced by objective measurement.

Misuses of Tests

Generally speaking, researchers are not studying or trying out and evaluating tests. They are studying other matters--problems, gains for compensatory programs, and the like. For the most part the tests are taken for granted as measuring instruments; in only a few cases are they questioned. That is undoubtedly why there are very few investigations of how well a test works--how valid it is--with specific differentiated groups. The published nationally standardized test is often accepted uncritically and/or simply used as the best available instrument for the purpose at hand.

Beyond the general acceptance of the test as "it," the search of the literature has uncovered some rather serious misuses of tests--using certain tests inappropriately, making comparisonsacross different tests, and reading into the test results more than the author and publisher intended. The Peabody Picture Vocabulary Test has been particularly misued. This easy-to-give test seems to be widely accepted as a good measure of general intelligence rather than offering an estimate (only) of verbal intelligence. It is frequently used with culturally deprived children with very limited vocabularies and the results compared with those of the norms group. Its use as a screening device is justified--nothing more.

Among other instances of misuse are these, which were written down as noted in reading the many studies abstracted for this report. The presence of a few such studies in this report is noted incidentally

-----Assuming that a test designed for gifted children of one age is suitable, then, for use with older children with limited backgrounds. (See Hunter Aptitude Scales study, p. 21)

-----More generally, assuming that a test constructed and standardized for children of a given age and/or school experience is equally valid for children of different ages and/or experience.

-----Changing some items and some credited answers, but applying the regular norms, especially with Puerto Rican and Mexican-American groups. (Noted in studies in preceding section)

-----Testing so early in preschool programs, in order to get a pretest base when improvement is to be measured, that test results cannot be valid. When a child has never handled pencil or crayon, never had a book or booklet and turned pages, never followed group directions, never worked steadily in a self-directed situation, then

a group test like the Metropolitan Readiness Test cannot be a valid measure. It does not measure what the test is designed to measure because test-taking is so new and unfamiliar. The resulting scores may be purely chance, or zero, although the children may have some degree of readiness.

Posttests after an interval of group experience and use of crayons, and so forth, can produce a more valid result. But to measure score gains from pre- to posttesting and ascribe them to the effectiveness of the program in bringing about improvement in the traits measured is not justifiable if no training for the pretesting has been given. (Several Headstart evaluations suffer from this flaw.)

-----Assuming that learning ability is measured by what has been learned, using the Peabody Picture Vocabulary Test or even the Stanford-Binet, with its heavy emphasis on vocabulary, or the Wechsler Intelligence Scale for Children, with children with limited backgrounds. The emphasis on evaluation in these early childhood programs should be on getting children ready to be taught. The emphasis should be on current achievement, rather than on "intelligence," in assigning them to learning groups.

-----Failing to separate reading and oral vocabulary in English from the appraisal of learning ability. Failure to use other than English-language tests for Mexican-American children, and then classifying low scoring pupils as mentally retarded, is a clear example. (Noted in preceding section)

-----Doing studies with very small numbers of students. In some studies, no tests of significance have been made and, if they had been, hardly any significant (meaningful) results could have been obtained because of the tremendous differences in score that would have been required. Many findings of "no significant difference" are attributable to the small numbers of cases involved.

-----Failing to follow through for two, three, four years, or more. The lack of longitudinal studies is distressing. It is little wonder that the longitudinal study of the culturally deprived in compensatory programs, being conducted under the auspices of Educational Testing Service for the U. S. Office of Education--from age three to grade 3--has been so widely hailed. There are no others like it.

-----Interpreting scores of individuals on short subtests when the reliability estimates, simply because of the length of the tests, make it impossible to trust the results of comparisons Comparison of means for groups on the same data would be quite permissible because group means are often quite reliable enough for such purposes.

-----Comparing reliability coefficients without reference to differences in range of scores.

-----Treating different measures of learning ability as though the results on them were comparable. Often, no attention is paid to what the test is measuring, i.e., to its content. Thus, the Goodenough-Harris Draw-a-Man and the Peabody Picture Vocabulary Test are often treated along

with the Stanford-Binet as though they were equivalent and similar measures. Results on group pencil-and-paper tests of mental ability cannot be treated as equivalent to the results from individual testing.

-----Attaching the same importance to predictive validity without intervention (in the form of compensatory training) as with it. When a minimum amount of intervention is used, predictive validity is an indicator of the usefulness of preliminary information; when substantial intervention is attempted, predictive validity is no longer subject to such simple interpretations. Successful intervention involves defeating predictions of failure.

Just as much of the research on ability grouping has failed to produce conclusive findings regarding the advantages (and the disadvantages) of such grouping, in like manner much of the research on the testing of the culturally limited has failed to produce conclusive findings regarding either the validity of the tests for the use being made of them or the validity of the interpretations of the test results for such students.

As long ago as 1964, Fishman et al. prepared a set of "Guidelines for Testing Minority Group Children." The reader may be referred to that source for a compact summary of the major issues.

The discussion in this document has taken particular account of their first two major points regarding the importance of any differences found in reliability and predictive validity when the same instruments are used to evaluate minority and majority group children. Notice has been taken at several junctures that (1) reliability of a test is often equally great for minority as for majority groups, and (2) predictive validity is often as high for minority or mixed groups as for majority groups. In fact, instances have been reported in which predictive equations based on majority groups overpredict the subsequent academic achievement of minority students, thereby "favoring" the minority groups at choice points such as college admission or ability group assignment.

The discussion proceeds farther, however, to consideration of factors that affect both measures taken at the inital point of prediction and the later "final" point of assessing achievement. It is here that doubt and confusion remain. Equally low effort and accomplishment at both points will contribute positively to predictive validity. Does this lack of effort on tests at both points, a failure to organize oneself for the ultimate in competitive effort, constitute a fundamental defect requiring remediation? Does modern life essentially require this competitive effort? If so, can it be learned? Meanwhile, what procedures can be adopted to keep these modifiable traits from unduly influencing initial measures? Can we turn to foreign students for a cue? Must we allow practically unlimited time for initially slow-paced children so they can take their time interpreting questions, reading and "translating" multiple-choice options, carrying through problem-solving operations?

Also, can we accept as a crucial goal of modern education the separation of essential objectives basic to success in school learning and later in employment from what have been considered marks of the educated person? If so, we may be able to foster affective development of minority children and thereby indirectly their cognitive development.

B. ALTERNATIVE STRATEGIES

The research into the procedures for the use of tests in grouping students for learning has provided limited information. This research has been described in earlier sections of this report as generally inconclusive, with the learning environment uncontrolled and the affective domain de-emphasized. There is real need for a well designed major program of longitudinal studies, including multi-variate and covariate analyses with consideration of the learning environment, in which the student's development is evaluated against criteria involving the cognitive, performance, and affective domains (Anderson, 1969). However, during the years required for such studies, certain helpful practices for the use of tests in the learning situation have been identified and can be described. The practices are concurred in by authorities from the fields of education and psychometrics.

Individualized Instruction

The purpose frequently stated for grouping children in learning situations is to provide for individual differences. In this subsection, selected procedures are discussed for test utilization and the realization of individualized instruction.

Perhaps, individualized instruction has as many definitions as there are "authorities" defining the term. Individualized instruction is herein thought of as a process of designing the curriculum for the individual (Goodlad, 1966; Rasmussen, 1968). In the process we would start by developing rapport with the student. As rapport is established the teacher initiates an effort to define the student's characteristics. If not initially, as soon as feasible, tests and measures should be utilized by a competent person to assist in the definition of the student's characteristics. As the student enters school, for example, the tests might well include individual intelligence tests and/or reading readiness measures.

After the teacher has established rapport with and gained a knowledge of the student, she is in a position to discuss objectives with the student. The objectives are mutually agreed upon and become those of the student. The curriculum content is selected by the teacher to support the student's objectives. The content includes relevant and realistic aspects of the cognitive, performance, and affective domains.

The student progresses at his rate in the mastery of the identified curricular content. It is emphasized that the student progresses at his rate to mastery. The mastery is normally determined in part, if not totally, by tests. The tests measure achievement and performance, and sample curricular content behaviors. The purpose of the testing is to establish mastery and readiness for the next curricular topic. In the event that the student has not mastered a given topic, he is not failed but continues to study the topic until mastery is obtained.

The procedures, materials, and methods used to guide the student in learning the content are individualized for the student (Glaser, 196?; Lindvall and Cox, 1969). In that the measures of cognitive processes and styles are in preliminary stages of development, they are not currently dependable for this purpose. Rather, the teacher should observe, both informally and systemically, the means whereby the student learns, and proceed to guide the student on a pragmatic basis.

Now that we have individualized instruction, is it possible to group students for learning? Four possible procedures are suggested. They are not exhaustive of all possible procedures. They are judged, in the light of the findings of the preceding sections, to be the most promising.

Heterogeneous Grouping

Heterogeneous grouping involves the bringing together of students who deviate extensively on a given variable. For example, in an elementary school social science class a topic for discussion might be the State of California. The student's knowledge of the state is the variable. Some student might have lived or visited in the state and observed a great amount of realistic information pertaining to the state. A group is formed consisting of those knowledgeable students and those desiring to learn about the state. In this instance we have an "ad hoc" heterogeneous group. The knowledgeable members have an opportunity to gain in leadership and communication skills through instruction of the others. The others, with guidance, are motivated to learn that which their peers know.

Heterogeneous grouping of this nature is practiced in the non-graded school. Children assigned in a non-graded school vary considerably in age, experience, and knowledge. The heterogeneity is planned so that the children can learn from each other.

Stratified Heterogeneous Grouping

The illustration just cited presents a clear case for the values of heterogeneous grouping. But let us consider another situation commonly faced in elementary schools in which it has been customary to teach classes of 30 children or so in self-contained classrooms where the 30 children stay with the same teacher in the same room for practically the entire day. Suppose we accept the criticism of those who argue for homogeneous ability grouping to reduce the span of achievement in each classroom, yet are even more attentive to the criticism of those who argue against homogeneous grouping of whole classrooms because of the stigma this places on those in the average and low groups while giving the high groups an unwholesome feeling of general superiority. Can these views both be accepted in a plan of organization of classrooms that has its own peculiar advantages? It has been done.

In Baltimore, a fundamental plan of organization recommended as an alternative that meets these requirements* may be called a plan of "stratified grouping." Under this plan, if three classes of 30 are to be made of 90 children ready to start fifth grade, the children would be ranked in order of excellence on some composite--say, a standardized test battery most recently given--and then be subdivided into nine groups of ten each. Teacher A would be given a class consisting of the highest or first ten, the fourth ten, and the seventh ten; Teacher B would have the second, fifth and eighth tens; Teacher C would then be given the third, the sixth and the ninth (lowest) tens.

Note the several merits of this scheme. First, there is no top or bottom section; the sections overlap, so invidious comparisons between groups are minimized. Second, each class has a narrower range than the full 90 have: Teacher A has the top ten, but none of the bottom 20; Teacher C has the bottom ten, but none of the top 20; Teacher B has neither the top nor the bottom ten. Third, teachers can give special attention where it is needed without feeling unable to meet the needs of the opposite extreme: Teacher A can give a little special attention to the top ten because the bottom 20 are not in the class; Teacher C can concentrate on the bottom ten, without fear of "losing" the top 20. Fourth, each class has leaders of appropriate capability to stimulate each other in a fair competitive way while giving leadership to lower groups; note particularly that in Teacher C's class, the top group is the third ten, a group that has probably always had to play second fiddle to some in the first or second ten. Finally, no teacher has to teach the bottom group of a homogeneous plan, that mixture of disruptive, leaderless children that lack motivation and capability and make teachers like homogeneous grouping, but equally dislike to teach the slow group.

Such a method of grouping is not offered as a complete answer by itself, but as a constructive step in the right direction. It is, moreover, compatible with other special teaching arrangements like team teaching, peer tutoring, and early education.

The history of heterogeneous grouping schemes is that they do not involve an additional expenditure of funds. Our third procedure is thought to involve additional funds, especially during the implementation phase. However, the additional gains in this third procedure are judged to show a favorable cost-effectiveness trade-off.

Team Teaching

The U. S. Office of Education has sponsored a number of efforts to develop specifications for new model elementary school systems. A total of ten (10) such models have been developed (Stauffer and Deal, 1969). Without exception each model, with numerous variations, has embraced the concepts of individualized instruction, mastery, and differentiated staff. The differentiated staff approach specifies various personnel categories for teachers such as aides, assistants, specialists, and the like (Allen, 1967). Each category has

*Elementary School Guide, Baltimore Public Schools, revised edition, 1967.

certain functions of prime responsibility. The team teaching staff is selected from these categories of teachers so as to satisfy the requirements of a given situation.

The team would normally contain or have readily available a specialist who would perform, or guide a competent teacher in, the diagnosis of the individual student. The specialist is trained in selecting and administering tests, interpreting test results, and defining appropriate programs of instruction. After the objectives and content are defined for the student, the task of guiding the student's learning is assigned among the team members as appropriate.

In a team, normally, there is a considerable number of staff members, say six or more, and a large class, say 100 or more. Thus, it is frequently found that a number of students have a need to learn the same tasks. Groups of such students are formed and assigned to a designated teacher for the purpose of learning the specific tasks. The grouping is informal, ad hoc, and of short duration. In a situation of this nature the students and teachers are paired with the task to be accomplished. Grouping in this manner promotes the effective utilization of personnel and resources, and increased learning by the individual, without the identified detrimental effect of homogeneous grouping.

Student Tutoring

Tutoring of children deficient in academic skills by older children has been widely adopted within compensatory education programs. Not surprisingly, those tutored show more than normal gains over a period of instruction. What is perhaps somewhat more surprising, when older children--themselves deficient in basic skills--are paid to tutor younger children who are deficient, the gains of the tutors outstrip by far the gains of the tutored!

Cloward (1967) reports a study in which children of junior high grade status who were two or more years retarded in reading, as measured by grade scores on a standardized reading test, were paid $1.25 per hour to tutor deficient fourth-grade children of similar ethnic background (Caucasian, Puerto Rican, Negro). The program was conducted over an academic year after the tutors had undertaken a period of preparation (also on paid time) for their teaching chores. The psychodynamics of the tutor growth is worth spelling out rather fully.

First, these older students who had experienced the constant role of failures pitied or deplored by their teachers were now being asked, nay even paid, to make a contribution to others. Second, in preparing for this work they had learned the basis of the old maxim "If you want to learn something, teach it." Third, they could see their pupils learn, as measured by daily response as well as by terminal test.

Specifically, using analyses of covariance to control for small initial differences in reading scores, Cloward found that 100 deficient readers in

fourth and fifth grade who were tutored for four hours a week for 26 weeks did reliably better than 79 control children at the end of that period, reversing somewhat the normal trend toward further retardation characteristic of their peers. Tests given five months apart showed average gains of 6 months by experimentals, 3½ months by controls. During the same period 77 tutors, who averaged 0.8 grades deficient at the start, gained reliably more than their 52 controls by 1.7 grades. Bearing in mind that grade score differences at high school level are magnified by the fact that the slope of the growth curve is decreasing, the adjusted mean difference at the end is slightly more than half a standard deviation on the score scale.

Early Childhood Education

At least since the 1930's, when the studies emanating from the Iowa Child Welfare Research Station (Stoddard, 1943) challenged the then accepted concept of the constancy of the IQ (Hunt, 1961) with evidence that substantial gains or losses in intellectual competence could be generated by the nature of early environmental stimulation of children, many parents from the upper socioeconomic classes have been sending their children to nursery schools. Beginning sometimes as early as age 2, these children have enjoyed intellectual stimulation in a supportive emotional climate and have emerged readier to participate in conventional schooling at age 5 or 6. In many such schools, priority has been given to affective development over intellectual stimulation. In others, however, intellectual stimulation has been an integral feature of this early education.

Currently, the debate rages about whether this early intellectual stimulation may be cast in a form that is best called early schooling, the earlier presentation of instructional stimulation ordinarily offered all comers at an approximately uniform starting point of age 6 in grade 1. What is best done at earlier ages is still moot, but experiments with children beginning at age 5 in kindergarten (McKee and Brzeinski (1966); Brzeinski et al., 1967; Fortson, 1969) show conclusively effective gains from planned early schooling in kindergarten. The Denver data reported by Brzeinski show that reliable gains from such early instruction in reading persist at least through grade 5, with some spread to related curriculum areas. An important condition is that gains achieved in kindergarten shall be consciously built upon in successive grades rather than being left to conventional programs for incidental forwarding; indeed, children placed in conventional classes with children beginning the learning of reading at age 6 in grade 1 soon slip from being recognized by their teachers as advanced at that point to becoming ones less challenged by the teaching of already learned skills and eventually being not at all advanced over their peers.

Implications of these and other findings for the enhancement of learning by disadvantaged groups would appear to be that the practice of beginning formal instruction at age 5 (with some imaginative adaptations) might well follow the established practice of the British infant school of beginning instruction for all children at this level.

A Note on Jensen and Other New Developments

Because of the widespread publicity achieved by the debate over an article entitled "How Much Can We Boost the IQ and Scholastic Achievement?" by Arthur R. Jensen in the Winter 1969 issue of the Harvard Educational Review, some readers may wonder at its relevance to the issue of ability grouping. Jensen suggests that some children learn better by association (rote memory), others by fitting new learning into a conceptual framework by higher mental processes, and that the whole matter of efficient learning styles is related to genetically determined "intelligence" in which certain ethnic groups are on the average considerably better endowed than others.

The reader is referred to the considerable bibliography of critical replies in subsequent issues of the Harvard Educational Review and elsewhere, listed at the end of this document. Suffice it here to quote from Cronbach's response and add our abbreviated critique.

Cronbach (1969) says in part:

> Professor Jensen is among the most capable of today's educational psychologists. His research is energetic and imaginative. In the present paper, an impressive example of his thoroughness, I am sure every reader has had my experience of encountering valuable information in areas where he thought himself au courant. Unfortunately, Dr. Jensen has girded himself for a holy war against "environmentalists" and his zeal leads him into over-statements and misstatements.

Despite the merits of Jensen's research remarked by Cronbach, and admitting the dubious propriety of some of the criticism addressed to Jensen for publishing data and argument that may be used for partisan ends, his presentation suffers from faults in at least five major respects:

1) He starts in journalistic style to proclaim a finding, rather than in professional style to build a convincing case.

2) Current brief and fragmented efforts at compensatory education show little effect, but it is too much to say compensatory education has failed. Efforts expended on short-term early education have produced modest gains in some instances; other experiments here and in other countries have succeeded (Brzeinski, 1967; Bloom, 1969). One might fairly add that no major effort comparable to the systematic discrimination of over three centuries against American blacks has even been attempted.

3) Traits with high heritability are often modifiable (Goldstein 1969).

4) Education's business is with a substantial modifiability. Even a correlation of .87 between monozygotic twins leaves 25% of the variance unaccounted for (Bloom 1969).

5) He closes on a note that suggests the likelihood of his model of distinctive learning styles for variously different children without clear evidence of the likely effectiveness of different teaching styles for classroom groups. Since disadvantagedness to

Jensen is an individual characteristic compounded of individual and group hereditary and environmental factors and their interactions, this can only imply responsiveness of teachers to all children with a variety of teaching styles rather than heavy dependence on one teaching style for children of each of the different learning styles. His discussion, moreover, leaves entirely out of consideration the teaching and learning that go on between children.

Other new proposals, like performance contracts and vouchering of funds to parents to let them "buy" their children's education from the best sources, are merely noted here. They are procedural rather than instructional variations. If used, it would remain for instruction to be designed as suggested here, or by more ingenious instructional plans; performance contracts and vouchering merely establish different contractual arrangements for authorizing instructional activity.

SUMMARY AND CONCLUDING REMARKS

After pointing out some of the pitfalls in the interpretation of tests used for grouping children with limited backgrounds and some of the efforts being made to provide better interpretative data, this document has been closed with a series of brief accountsof alternative strategies to ability grouping. These illustrations by no means exhaust the possibilities, but they constitute a set of mutually compatible strategies each of which has separate merit. Heterogeneous grouping promotes communication and peer teaching. Stratified heterogeneous grouping furthers these same goals while reducing the extreme variations in a class that complicate group instruction. Team teaching permits flexible grouping to achieve individual learning objectives. Student tutoring promotes learning by the tutors as well as by the tutored, a circumstance also furthered by stratified grouping. Early childhood education, at least from kindergarten at age 5, can undergird a persistent gain in mastery of fundamentals. Taken together, these alternative strategies constitute a constructive challenge to the unrealized advantages and actual deleterious effects of ability grouping in the areas of scholastic achievement, affective development, and the ethnic and socioeconomic separation (isolation, deprivation) of children.

References

Abramson, Theodore. "The Influence of Examiner Race on First-Grade and Kindergarten Subjects' Peabody Picture Vocabulary Test Scores," Journal of Educational Measurement, 6:241-46, Winter 1969.

Allen, D. W. "Differential Teaching Staff." Interim seminar paper presented at Stanford University, Stanford, California, March 1967.

Alzobaie, Abdul Jabil, Metfessel, Newton S. and Michael, William B. "Alternative Approaches to Assessing the Intellectual Abilities of Youth from a Culture of Poverty," Educational and Psychological Measurement, 28:449-55, Summer 1968.

Anastasi, Anne. "Culture-Fair Testing." Educational Horizons, 48:26-30, Fall 1964.

Anderson, Scarvia B. "The ETS-OEO Longitudinal Study of Disadvantaged Children." In Untangling the Tangled Web of Education, a special symposium sponsored by the National Council on Measurement in Education. Princeton, New Jersey: Educational Testing Service, 1969. pp. 27-38.

Baltimore Public Schools. Elementary School Guide, revised edition. Baltimore, Maryland, 1967.

Beidler, Anne E. "The Effects of the Peabody Language Development Kits on the Intelligence, Reading, Listening, and Writing of Disadvantaged Children in the Primary Grades," Bethlehem, Pennsylvania: Legigh University, 1968. Abstract: Dissertation Abstracts 29:3760-A, May 1969.

Boney, J. Don. "Predicting the Academic Achievement of Secondary School Negro Students." Personnel and Guidance Journal, 44:Part 2, 700-03, March 1966.

Bradley, Nolen E. "The Negro Teacher-Graduate Student: Factors Relative to Performance in Predominantly White State Colleges and Universities in Tennessee," Journal of Negro Education, 36:15-23, Winter 1967

Bloom, Benjamin S Letter to the Editor. Harvard Educational Review, 39:419-21, Spring 1969.

Brazziel, W. F. and Terrell, Mary. "An Experiment in the Development of Readiness in a Culturally Disadvantaged Group of First Grade Children." Journal of Negro Education, 31:4-7, Winter 1962.

Brzeinski, Joseph E., Harrison, M. Lucile, and McKee, Paul "Should Johnny Read in Kindergarten?" NEA Journal, 56 (3):23-25, March 1967.

Buchanan, Richard G. "The Effect of Cultural Deprivation on Test-Taking Approach as Indicated by Response Style to Multiple-Choice Questions." New York: New York University, 1968. Abstract: Dissertation Abstracts 29:2994-A, March 1969.

California Elementary School Administrators Association. "A Study of Socio-Economic Status and School Achievement." 1962 (c/o California Teachers Association, 1705 Murchison Dr., Burlingame, California, 94010)

Campbell, Joel. "Testing of Culturally Deprived Groups." Unpublished Research Bulletin. RB-64-34. Princeton, New Jersey: Educational Testing Service, June 1964.

Carroll, John B. "Possible Directions in which College Board Tests of Abilities and Learning Capacities Might Be Developed." In Report of the Commission on Tests II Briefs. New York: College Entrance Examination Board, 1970. pp. 1-12.

Carver, Ronald P. "An Experiment that Failed: Designing an Aural Aptitude Test for Negroes." College Board Review, 70:10-14, Winter 1968-69.

Carver, Ronald P. "The Questionable Uniqueness of a Newly Developed Listening Test." American Educational Research Journal, 5:728-30, November 1968.

Carver, Ronald P. "Use of a Recently Developed Listening Comprehension Test to Investigate the Effect of Disadvantagement upon Verbal Proficiency." American Educational Research Journal, 6:263-70, March 1969.

Cattell, Raymond B. "Theory of Fluid and Crystalized Intelligence: A Critical Experiment." Journal of Educational Psychology, 54:1-22, February 1963.

Chandler, John T. and Plakos. "Spanish-Speaking Pupils Classified as Educable Mentally Retarded. California State Department of Education: Mexican American Education Research Project." Integrated Education, 7(6):28-33, November-December 1969.

Cleary, T. Anne. "Test Bias: Prediction of Grades of Negro and White Students in Integrated Colleges." Journal of Educational Measurement, 5:115-24, Summer 1968.

Cleary, T. Anne and Hilton, Thomas L. "An Investigation of Item Bias" Educational and Psychological Measurement, 28:61-75, Spring 1968.

Cloward, Robert D. "Studies in Tutoring." Journal of Experimental Education, 36:14-25, Fall 1967.

College Entrance Examination Board. THE COLLEGE BOARD ADMISSIONS TESTING PROGRAM: A Technical Report on Research and Development Activities Relating to the Scholastic Aptitude Test and Achievement Tests. William H. Angoff, editor. New York: College Entrance Examination Board (in press).

Commission on Tests, Tiedeman, David V., chairman. Report of the Commission on Tests: I. Righting the Balance. II. Briefs. New York: College Entrance Examination Board, 1970.

Cronbach, L. J. "Heredity, Environment and Educational Policy." Harvard Educational Review, 39:338-47, Spring 1969.

Davis, Allison, Chairman; Eells, Kenneth; Havighurst, Robert J.; Herrick, Virgil E; and Tyler, Ralph W. Intelligence and Cultural Differences Chicago: University of Chicago Press, 1951

Davis, Elred Dennis. "Test Performance in Communication Skills of Pupils Attending Schools in Disadvantaged Areas of Knoxville (1965-67)." Knoxville University of Tennessee, 1968. Abstract: Dissertation Abstracts 29:3868 A, May 1969.

Davis, O. L., Jr., and Personke, Carl R., Jr. "Effects of Administering the Metropolitan Readiness Test in English and Spanish to Spanish-speaking School Entrants." Journal of Educational Measurement, 5:231-34, Fall 1968.

Dowd, Gerold John. "Sex and Race Differences in the Effectiveness of Various Composite Predictors of Initial Reading Success and the Relationship of Children's Self-Perceptions to Initial Reading Success." Jamaica, New York: St. John's University, 1968. Abstract: Dissertation Abstracts 29:2999-A, March 1969.

Dreger, R. M. and Miller, D. S. "Comparative Psychological Studies of Negroes and Whites in the United States: 1959-1965." Psychological Bulletin (Monograph Supplement, 70, No. 3, Part 2) 1968.

Dubin, Jerry A. and Osburn, Hobart. "Speed and Practice: Effects on Negro and White Test Performance." Journal of Applied Psychology, 53:19-23, February 1969.

Eagle, Norman and Harris, Anna S. "Interaction of Race and Test on Reading Performance Scores." Journal of Educational Measurement, 6:131-35, Fall 1969

Eells, Kenneth; Davis, Allison, chairman; Havighurst, Robert J.; Herrick, Virgil E; and Tyler, Ralph W. Intelligence and Cultural Differences, A Study of Cultural Learning and Problem-solving. Chicago: University of Chicago Press, 1951.

Feldmann, Shirley C. "Predicting Early Success." Paper read at conference of International Reading Association in Detroit, May 1965.

Fishman, Joshua, chairman; Deutsch, Martin; Kogan, Leonard; North, Robert; and Whiteman, Martin. "Guidelines for Testing Minority Group Children." Prepared by a Work Group of the Society for the Psychological Study of Social Issues. Journal of Social Issues, 20:129-45, April 1964.

Flaugher, Ronald L. "Testing Practices, Minority Groups, and Higher Education: A Review and Discussion of the Research." unpublished Research Bulletin, RB-70-41. Princeton, New Jersey: Educational Testing Service, June 1970.

Fortson, Laura R. "A Creative Aesthetic Approach to Readiness and Beginning Reading and Mathematics in Kindergarten." Athens: University of Georgia, 1969. Abstract: Dissertation Abstracts 30:5346-A, June 1970.

Glaser, Robert. "The Design of Instruction." In Goodlad, John I., editor, The Changing American School, Sixty-fifth Yearbook of the National Society for the Study of Education, Part II, pp. 215-42. Chicago: University of Chicago Press, 1966.

Goldstein, Leo S.; Coller, Alan R.; Dill, John; and Tilis, Howard S. "The Effect of a Special Curriculum for Disadvantaged Children on Test-Retest Reliabilities of Three Standardized Instruments." Journal of Educational Measurement, 7:171-74, Fall 1970.

Goodlad, John I. "Diagnosis and Prescription in Educational Practice." In New Approaches to Individualizing Instruction, a report of a conference in May, 1965, to mark the dedication of Ben D. Wood Hall. Princeton, New Jersey: Educational Testing Service, 1966, pp. 27-37.

Green, Robert L. and Hoffman, Louis J. "A Case Study of the Effects of Educational Deprivation on Southern Rural Negro Children." Journal of Negro Education, 34:327-41, Summer 1965.

Harris, Albert J. and Lovinger, Robert J. "Longitudinal Measures of the Intelligence of Disadvantaged Negro Adolescents." School Review, 76:60-66, March 1968.

Hills, J. R.; Klock, J.C.; and Lewis, S. Freshman Norms for the University System of Georgia, 1960-62. Atlanta, Georgia Regents of the University System, 1963.

Hughes, Robert B. and Lessler, Ken. "A Comparison of WISC and Peabody Scores of Negro and White Rural School Children." American Journal of Mental Deficiency, 69:877-80, May 1965.

Hunt, J. M. Intelligence and Experience. New York: Ronald Press, 1961.

Jensen, Arthur R. "How Much Can We Boost the IQ and Scholastic Achievement?" Harvard Educational Review, 39:1-123, Winter 1969.

Karabinus, Robert A. and Hurt, Maure, Jr. "The Van Alstyne Picture Vocabulary Test Used with Six-year-old Mexican-American Children." Educational and Psychological Measurement, 29:535-39, Winter 1969.

Kariger, Roger H. "The Relationship of Lane Grouping to the Socioeconomic Status of the Parents of Seventh-Grade Pupils in Three Junior High Schools." East Lansing, Michigan: Michigan State University, 1962. Abstract: Dissertation Abstracts 23:4586, June 1963.

Kendrick, S. A. and Thomas, Charles L. "Transition from School to College" in Education for the Disadvantaged, Edmund Gordon, issue ed. Review of Educational Research, 40:151-79, February 1970.

Kennedy, Wallace A.; VanDeRiet, Vernon; and White, James C., Jr., "A Normative Sample of Intelligence and Achievement of Negro Elementary School Children in the Southeastern United States." Monograph of the Society for Research in Child Development, 28:6, No. 90. Lafayette, Indiana: Child Development Publications for the Society for Research in Child Development, 1963.

Kimball, William L. "Parent and Family Influences on Academic Achievement Among Mexican-American Students. Los Angeles: University of California, Los Angeles. Abstract: Dissertation Abstracts 29:1965-A, December 1968.

Kneif, L. M. and Stroud, J. B. "Intercorrelations among Various Intelligence, Achievement, and Social Class Scores." Journal of Educational Psychology, 50:117-20, June 1959.

Lambert, Nadine M. "The Present Status of the Culture Fair Testing Movement." Psychology in the Schools, 1:318-30, July 1964.

Lesser, Gerald S.; Fifer, Gordon; and Clark, Donald H. "Mental Abilities of Children from Different Social-Class and Cultural Groups." Monograph of the Society for Research in Child Development, 30:4, No. 102. Chicago: University of Chicago Press, 1965.

Lindvall, C.M. and Cox, R.C. "The Role of Evaluation in Programs for Individualized Instruction." In Tyler, Ralph W., editor, Educational Evaluation: New Roles, New Means, Sixty-eighth Yearbook of the National Society for the Study of Education, Part II. Chicago: University of Chicago Press, 1969, pp. 156-88.

LoMonaco, Leon John. "Response Levels of Disadvantaged Ninth-Grade Boys to Both Standard and Oral-Visual Administrations of Two Vocationally Relevant Instruments." New York: New York University, 1968. Abstract: Dissertation Abstracts 29:3004-A, March 1969.

Lorge, Irving. "Difference or Bias in Tests of Intelligence." Paper presented at 1952 Invitational Testing Conference of Educational Testing Service. In Anastasi, Anne, editor, Testing Problems in Perspective. Washington, D.C.: American Council on Education, 1966, pp. 465-71.

Lucas, Charles M. "Survey of the Literature Relating to the Effects of Cultural Background on Aptitude Test Scores." Unpublished Research Bulletin, RB-53-13. Princeton, New Jersey: Educational Testing Service, June 30, 1953.

Machover, Solomon. "Cultural and Racial Variations in Patterns of Intellect; Performance of Negro and White Criminals on the Bellevue Adult Intelligence Scale." Teachers College Contributions to Education. No. 875, 1943; also Teachers College Record, 45:52-54, October 1943 (summary).

McKee, Paul R. and Brzeinski, J.E., The Effectiveness of Teaching Reading in Kindergarten. Cooperative Research Project No. 5-0371. Denver, Colorado: Denver Public Schools, 1966.

Metfessel, Newton S. "An Investigation of Attitudinal and Creativity Factors Related to Achieving and Nonachieving Culturally Disadvantaged Youth." Project Potential, USOE Cooperative Research Project No. 2615. Los Angeles, California: University of Southern California, 1965.

Mitchell, Blythe C. "Predictive Validity of the Metropolitan Readiness Tests and the Murphy-Durrell Reading Readiness Analysis for White and Negro Pupils." Educational and Psychological Measurement. 27:1047-54, Winter 1967, Part II.

Morper, Jack. "An Investigation of the Relationship of Certain Predictive Variables and Academic Achievement of Spanish-American and Anglo Pupils in Junior High School." Stillwater: Oklahoma State University, 1966. Abstract: Dissertation Abstracts 27:4051-A, June 1967.

Munday, Leo. "Predicting College Grades in Predominantly Negro Colleges." Journal of Educational Measurement, 2:157-60, December 1965.

Orr, David B. and Graham, Warren R. "Development of a Listening Comprehension Test to Identify Educational Potential Among Disadvantaged High School Students." American Educational Research Journal, 5:167-80, March 1968.

Palomares, Uvaldo Hill and Johnson, Laverne C. "Evaluation of Mexican American Pupils for Educable Mentally Retarded Classes." California Education, 3(8):27-29, April 1966.

Pelosi, John William. "A Study of the Effects of Examiner Race, Sex, and Style on Test Responses of Negro Examinees." Syracuse New York: Syracuse University, 1968. Abstract: Dissertation Abstracts 29:4105-A, May 1969.

Personke, Carl R., Jr., and Davis, O.L., Jr. "Predictive Validity of English and Spanish Versions of a Readiness Test." The Elementary School Journal 70:79-85, November 1969.

Pettigrew, Thomas F. A Profile of the Negro American. Princeton, New Jersey: D. Van Nostrand Company, Inc., 1964.

Rasmussen, L.V. Meeting the Critics' Demands for Quality Education--through Individualized Instruction. A special report. Washington, D.C.: The National Laboratory for the Advancement of Education, 1968, pp. 12-15.

Roberts, S. O. Studies in Identification of College Potential. Nashville, Tennessee: Fisk University, Department of Psychology, 1962. (Mimeographed)

Roberts, S. Oliver: Crump, E.P.; Dickerson, Ann E.; and Horton, C.P. "Longitudinal Performance of Negro-American Children at Five and Ten Years on the Stanford Binet." Paper read at annual meeting of American Psychological Association, September 1965.

Santos, Beatriz N. "Special Achievement Testing Needs of the Educationally Disadvantaged." Iowa City: University of Iowa, 1967. Abstract: Dissertation Abstracts 28:2567-A, January 1968.

Shuey, Audrey. The Testing of Negro Intelligence, 2nd Edition. New York: Social Science Press, 1966.

Society for the Psychological Study of Social Issues. "Guidelines for Testing Minority Group Children." Prepared by a Work Group, Joshua Fishman, chairman. Journal of Social Issues, 20:129-45, April 1964.

Stanley, J.C. and Porter, A.C. "Correlation of Scholastic Aptitude Test Scores with College Grades for Negroes versus Whites." Journal of Educational Measurement, 4:199-218, Winter 1967.

Stauffer, A. J. and Deal, T.N., editors. "Teacher Evaluation Models." Journal of Research and Development in Education, 2(3):3-135, Spring 1969.

Stoddard, George D. The Meaning of Intelligence. New York: Macmillan Co., 1943.

Temp, George. "Test Bias: Validity of the Scholastic Aptitude Test for Blacks and Whites in Thirteen Integrated Institutions." Princeton, New Jersey: Educational Testing Service (in preparation).

Turnbull, William W. "Influence of Cultural Background on Predictive Test Scores." Paper presented at 1949 Invitational Testing Conference of Educational Testing Service. In Anastasi, Anne, editor, Testing Problems in Perspective. Washington, D.C.: American Council on Education, 1966, pp. 458-64.

Weiner, Max and Feldmann, Shirley. "Validation Studies of a Reading Prognosis Test for Children of Lower and Middle Socio-economic Status." Educational and Psychological Measurement, 23:807-14, Winter 1963.

Wilson, Kenneth R. "Black Students Entering CRC Colleges: Their Characteristics and Their First-Year Academic Performance." Research Memorandum 69-1. Poughkeepsie, New York: College Research Center, April 15, 1969. Mimeographed.

Additional Commentary on Jensen's Thesis

Anastasiow, N. J. "Educational Relevance and Jensen's Conclusions." Phi Delta Kappan, 51:32-5, Summer 1969.

Bereiter, Carl. "The Future of Individual Differences." Harvard Educational Review, 39:310-18, Spring 1969.

Brazziel, William F. "A Letter from the South." Harvard Educational Review, 39:348-56, Spring 1969.

Brazziel, W. F. "Perspective on the Jensen Affair." Childhood Education, 46:371-2, April 1970.

Brazziel, William F. "Symposium: The Jensen Controversy. I. Beyond the Sound and Fury." Measurement and Evaluation in Guidance, 3:7-9, Spring 1970.

Brazziel, William F., Brown, Frederick G., and Cameron, Howard. "Symposium: The Jensen Controversy." Measurement and Evaluation in Guidance, 3:7-24, Spring 1970.

Brown, Frederick G. "Symposium: The Jensen Controvery. III. Review of the Past, Focus for the Future." Measurement and Evaluation in Guidance, 3:18-24, Spring 1970.

Cameron, Howard. "Symposium: The Jensen Controversy. II. Cultural Myopia." Measurement and Evaluation in Guidance, 3:10-17, Spring1970.

Comer, J. P. "Research and the Black Backlash." American Journal of Orthopsychiatry, 40:8-11, January 1970.

Crow, James F. "Genetic Theories and Influences: Comments on the Value of Diversity," Harvard Educational Review, 39:301-09, Spring 1969.

Deutsch, M. "Happenings on the Way Back to the Forum." Harvard Educational Review, 39:523-57, Summer 1969.

Elkind, David. "Piagetian and Psychometric Conceptions of Intelligence." Harvard Educational Review, 39:319-37, Spring 1969.

Fehr, F. S. "Critique of Hereditarian Accounts of Intelligence and Contrary Findings." Harvard Educational Review, 39:571-80, Summer 1969.

Goldstein, Allen C. "A Flaw in Jensen's Use of Heritability Data." IRCD Bulletin, 5:4:7-9, Fall 1969.

Gordon, Edmund W. "Education, Ethnicity, Genetics, and Intelligence." IRCD Bulletin, 5:4:1, 2, 13, Fall 1969.

Gruber, Howard E. "Psychologists Comment on Current IQ Controversy: Heredity versus Environment." IRCD Bulletin, 5:4:6, Fall 1969.

Hirsch, Jerry. "Behavior-Genetic Analysis and Its Biosocial Consequence." IRCD Bulletin, 5:4:3-4, Fall 1969.

Hudson, L. "Nature, Nurture: Racialist Comeback?" Times Educational Supplement, 2824:33, July 4, 1969.

Humphreys, L. G. and Dachler, H. P. "Jensen's Theory of Intelligence." With Reply by Jensen and rejoinder by authors. Journal of Educational Psychology, 60:419-33, December 1969.

Hunt, J. McV. "Has Compensatory Education Failed: Has It Been Attempted?" Harvard Educational Review, 39:278-300, Spring 1969.

Jensen, Arthur R. "Jensen's Theory of Intelligence." Journal of Educational Psychology, 60:427-31, December 1969.

Jensen, Arthur R. "Reducing the Heredity - Environmental Uncertainty: A Reply." Harvard Educational Review, 39:449-83, Summer 1969.

Kagan, Jerome S. "Inadequate Evidence and Illogical Conclusions." Harvard Educational Review, 39:274-77, Spring 1969.

Light, R. J. and Smith, P. V. Social Allocation Models of Intelligence." Harvard Educational Review, 39:484-510, Summer 1969.

Scriven, Michael. "The Values of the Academy (Moral Issues for American Education and Educational Research Arising from the Jensen Case). Review of Educational Research, 40:541-49, October 1970.

Stinchcombe, A. L. "Environment: the Cumulation of Effects." Harvard Educational Review, 39:511-22, Summer 1969.

Voyat, Gilbert. "IQ: God-given or Man-made?" Education Digest 35:1-4, October 1969.

Zach, L. "I.Q. Test: Does It Make Black Children Unequal?" School Review, 78:249-58, February 1970.

TEST REFERENCES

Titles	Publisher	Page (s)
AMERICAN COLLEGE TESTS (ACT)	American College Testing Program	29, 30
CALIFORNIA ACHIEVEMENT TESTS	California Test Bureau	23, 25
CALIFORNIA TESTS OF MENTAL MATURITY	California Test Bureau	29
CATTELL CULTURE-FAIR INTELLIGENCE TEST	Bobbs-Merrill Co., Inc.	28
CLYMER-BARRETT PREREADING BATTERY	Personnel Press, Inc.	6
COLLEGE BOARD ACHIEVEMENT TESTS	College Entrance Examination Board	29
COLUMBIA MENTAL MATURITY SCALE	Harcourt Brace Jovanovich, Inc.	21
COOPERATIVE PRIMARY TESTS	Educational Testing Service	22
DAVIS-EELLS GAMES	Harcourt Brace Jovanovich, Inc.	25
DIFFERENTIAL APTITUDE TESTS	Psychological Corporation	29
EMPLOYEE APTITUDE SURVEY	Psychological Services, Inc.	20
ENVIRONMENTAL PROCESS SCALE	*	6
GATES PRIMARY READING TESTS: WORD RECOGNITION, SENTENCE READING, PARAGRAPH READING	Teachers College Press	16, 22
HOLLINGSHEAD AND REDLICH INDEX	*	21
HOLLINGSHEAD TWO-FACTOR INDEX	*	23
HUNTER APTITUDE SCALE	*	21
INDIVIDUAL TESTS OF CREATIVITY	*	32
INVENTORY OF SELF-APPRAISAL	*	32
IOWA TESTS OF BASIC SKILLS	Houghton-Mifflin Co.	25, 26
IPAT CULTURE FAIR INTELLIGENCE TEST (SEE CATTELL CULTURE-FAIR INTELLIGENCE TEST.)	Institute for Personality & Ability Testing Bobbs-Merrill Co., Inc.	19
KUDER PREFERENCE RECORD-VOCATIONAL	Science Research Associates, Inc.	27
LEE-CLARK READING READINESS TEST	California Test Bureau	22
LIFE-PLANNING QUESTIONNAIRE-MODIFIED	*	27
LORGE-THORNDIKE INTELLIGENCE TESTS	Houghton-Mifflin Co.	15, 25, 28, 31
MEANING OF WORDS INVENTORY	*	32
METROPOLITAN ACHIEVEMENT TESTS	Harcourt Brace Jovanovich, Inc.	16, 25, 26, 27, 30, 31
METROPOLITAN READINESS TESTS	Harcourt Brace Jovanovich, Inc.	7, 8, 9, 10, 11, 12, 22, 30, 31
METROPOLITAN READING TEST (SEE METROPOLITAN ACHIEVEMENT TESTS)	Harcourt Brace Jovanovich, Inc. Harcourt Brace Jovanovich, Inc.	16, 27

*No publishers identified.

Titles	Publisher	Page (s)
MURPHY-DURRELL READING READINESS ANALYSIS	Harcourt Brace Jovanovich, Inc.	9, 10
ORR-GRAHAM LISTENING TEST	American Institutes of Research	16, 17
OTIS-LENNON MENTAL ABILITY TEST	Harcourt Brace Jovanovich, Inc.	13, 14, 22
OTIS QUICK-SCORING MENTAL ABILITY TEST-ALPHA	Harcourt Brace Jovanovich, Inc.	28
OTIS QUICK-SCORING TEST OF MENTAL ABILITY-BETA	Harcourt Brace Jovanovich, Inc.	28
PEABODY PICTURE VOCABULARY TEST	American Guidance Service, Inc.	20, 21, 24
PINTNER-CUNNINGHAM PRIMARY TEST	Harcourt Brace Jovanovich, Inc.	28
PINTNER GENERAL ABILITY TEST	Harcourt Brace Jovanovich, Inc.	28
PRELIMINARY SCHOLASTIC APTITUDE TEST	College Entrance Examination Board	14
PRESCHOOL INVENTORY	Educational Testing Service	6
PURDUE PEGBOARD	Science Research Associates, Inc.	
Raven's PROGRESSIVE MATRICIES	Psychological Corporation (U. S. Distribution)	25
READING PROGNOSIS TEST	Institute of Developmental Studies	16
SCHOLASTIC APTITUDE TEST	College Entrance Examination Board	14, 15, 29
SCHOOL AND COLLEGE ABILITY TEST (SCAT)	Educational Testing Service	17, 23, 29 31
SEQUENTIAL TESTS OF EDUCATIONAL PROGRESS (STEP)	Educational Testing Service	17, 23
STANFORD ACHIEVEMENT TEST	Harcourt Brace Jovanovich, Inc.	9, 10, 11, 12, 27
STANFORD-BINET INTELLIGENCE SCALE	Houghton-Mifflin Co.	23, 24, 27, 31
STEP LISTENING (SEE SEQUENTIAL TESTS OF EDUCATIONAL PROGRESS)	Educational Testing Service	17
STEP READING (SEE SEQUENTIAL TESTS OF EDUCATIONAL PROGRESS)	Educational Testing Service	17
TEST OF ACADEMIC PERFORMANCE-READING	*	28
U-SCALE	*	22
VAN ALSTYNE PICTURE VOCABULARY TEST	Harcourt Brace Jovanovich, Inc.	22, 31
Warner INDEX OF SOCIAL CLASS	*	25
Warner INDEX OF STATUS CHARACTERISTICS	*	25
WECHSLER ADULT INTELLIGENCE SCALE (WAIS)	Psychological Corporation	19
WECHSLER INTELLIGENCE SCALE FOR CHILDREN (WISC)	Psychological Corporation	24, 28, 31, 32
[illegible] I HOME ENVIRONMENT SCALE	*	6

*No publishers identified.