DOCUMENT RESUME

ED 113 371 TM 004 868

AUTHOR Ratteray, Joan D.
TITLE The Testing of Cultural Groups. A Paradigmatic Analysis of the Literature on Testing and a Proposition.
INSTITUTION Rand Corp., Santa Monica, Calif.
REPORT NO P-5403
PUB DATE Nov 74
NOTE 65p.
AVAILABLE FROM Rand Corporation, Santa Monica, Calif. 90406 ($5.00)

EDRS PRICE MF-$0.76 HC-$3.32 Plus Postage
DESCRIPTORS Compensatory Education; Criterion Referenced Tests; Cultural Differences; *Culture Free Tests; Educational Alternatives; Educational History; Educational Legislation; Elementary Secondary Education; *Ethnic Groups; Literature Reviews; Minority Groups; *Nature Nurture Controversy; Political Issues; Predictive Validity; Racial Differences; Standardized Tests; *Test Bias; *Testing; Testing Problems; Test Validity

ABSTRACT

Numerous strategies have been used throughout the years to test cultural groups. This paper grew out of the need to find and use standardized tests that would depict accurately the performance of various cultural groups in America. In order to make judgments about performance, it is wise to examine the theoretical structure from which most of the existing tests were developed. Accordingly, the paper traces the development of the various strategies and theoretical structures, explaining why they have met with limited success. Through a paradigmatic analysis of the literature, it identifies the existing testing paradigm as a monocultural one, and it relates the various efforts to produce a culture-fair test. The paradigmatic analysis is extended to encompass a proposition, based upon the coalescence of the scientific (theoretical and measurement) and policy contexts. The analysis suggests a procedure by which tests can be developed and/or evaluated if they are to depict accurately the performance of various cultural groups. (Author)

THE TESTING OF CULTURAL GROUPS

A Paradigmatic Analysis of the Literature on Testing and a Proposition

Joan D. Ratteray

November 1974

P-5403

The Rand Paper Series

The Rand Corporation
Santa Monica, California 90406

CONTENTS

GENERAL INTRODUCTION

Numerous strategies have been used throughout the years to test cultural groups. This paper grew out of the need to find and use standardized tests that would depict accurately the performance of various cultural groups in America.

In order to make judgments about performance, it is wise to examine the theoretical structure from which most of the existing tests were developed. Accordingly, the paper traces the development of the various strategies and theoretical structures, explaining why they have met with limited success. Through a paradigmatic analysis of the literature, it identifies the existing testing paradigm as a monocultural one, and it relates the various efforts to produce a culture-fair test.

The paradigmatic analysis is extended to encompass a proposition, based upon the coalescence of the scientific (theoretical and measurement) and policy contexts. The analysis suggests a procedure by which tests can be developed and/or evaluated if they are to depict accurately the performance of various cultural groups.

PART I

THE MONOCULTURAL TRADITION OF TESTING

In this paper, the testing tradition has been subjected to a paradigmatic analysis, which is the examination of the nature and dimensions of a testing paradigm.* It is shown that the study of cultural groups was accomplished by the application of the existing testing process.

A: DEVELOPMENT OF THE EXISTING PARADIGM

NATURE OF THE EXISTING TESTING PARADIGM

The contemporary testing paradigm is monocultural in nature, that is, it has relevance primarily for one dominant cultural group. It is born of a tradition that has elevated testing to a position of prestige and influence in the American way of life. It has been accomplished through several historical, social and economic events. Such events have included the growth of the melting pot concept and the emergence of the criteria for cultural group separation.

THE GROWTH OF THE "MELTING POT"

The concept of the "melting pot" was greatly responsible for achieving the cultural homogeneity needed for certain groups to be assimilated into the political, legal and social developments of American life.

Historical precedents can explain the emergence of such a timely concept. The Anglo-Saxon cultural groups from northern and western Europe were heavily represented in earlier populations who immigrated to this country. Eventually, the magnitude and the diversity of cultural groups were extended to populations from southern and eastern Europe.

---

*A paradigm consists of dimensions or "sets" that have common postulates and uniformly accepted meanings that have been attributed to those postulates.

Millions of immigrants brought strong tendencies toward cultural diversity which were incorporated into the American heritage. There was a need, however, to solidify these differing views into a common frame of reference if the American culture were to withstand the impact of the cross-cultural varieties of poverty, wealth, geography, religiosity and language. The structure for this frame of reference had been set by Anglo-Saxon cultural tradition, for it was the earlier "native" colonial Americans that determined the development of the American political, social and legal systems.* These systems were responsible for a massive standardization of the country's expanding institutions, and for motivating and syncronizing the specialization of those institutions, whereby there could be a superficial disappearance of external cultural differences. The absence of different social treatment for the groups lessened the impact of the cultural differences and permitted the groups to co-exist peacefully. The unity of the groups was further solidified in the monocultural tradition by the adoption of an American version of English speech.

The maintenance of the monocultural tradition has resulted in the concept of the "melting pot," a perspective that remained a myth for non-European cultural groups. The melting pot did not at that time become a reality even for some European immigrants who were non-English speaking.** Quota laws for Europeans after World War I were more directly controlled by the Federal Government rather than by the individual states. The immigration of groups from southern and eastern Europe was limited by law, but not in practice.† At first, neither the non-European nor the European groups were allowed to blend into the politically and economically unified dominant group. Eventually, however, the latter did merge.

---

*Carl Wittke, "Historical Background: Immigration Policy Prior to World War I," found in Benjamin M. Ziegler (ed.) *Immigration: An American Dilemma*, D.C. Heath and Co., Boston, 1953, pp. 1-10.

**Wittke, *loc. cit.*, p. 2.

†Ziegler, *Immigration*, Boston: Heath, 1953.

The reason for the difference in treatment between the two groups was that the non-Europeans were subjected to the enactment of special immigration laws, a tradition of slavery, and the colonial occupation of Indian land by the Europeans.

In a cultural sense, these non-European groups have been separated from the "dominant" European group by the "entry status"* accorded them as they became a part of the American scene. This primary distinction of entry status separates the "dominant" and "minority" cultures, the latter being defined as Blacks, Chicanos, Native-Americans and Oriental/Asiatics, and present-day population statistics support this distinction. An interesting corollary is the proportion of these groups which is below the official poverty level in the United States.**

Because cultural heterogeneity and low socio-economic statuses have compounded the assimilation process into the "dominant" American culture for these groups, a type of cultural disenfranchisement of these minority groups has resulted.

DIMENSIONS OF THE EXISTING PARADIGM

Scientific history has caused three dimensions or "sets"† to converge in the testing process: the theoretical, methodological and functional. The assumptions underlying these dimensions were predicted upon and/or inspired by the European tradition and were incorporated into the rapidly changing American society.

---

*Such a status can be exemplified by African slaves, Mexican-American migrant farm workers, Puerto Rican immigrants from protectorates in colonial era, Native Americans who seek initiation into the mainstream from the reservations, Japanese Americans in West coast concentration camps during World War II, Chinese Americans in defined upper social and economic statuses rather than status as railroad laborers (Circa 1866).

**Decennial Census data 1970: Of 87.5% of the "dominant" group 5.3% is below poverty line; Minority cultural groups: 11.1% (Blacks, with 29.9% below poverty); Other groups make up 1.4% of the populations (each separate group make up less than .1% by itself). The proportion below poverty line include: Filipino 11.5%; American Indians 33.3%; Spanish surname 23.4%; Spanish origin 21.1%; Japanese 6.4%; Chinese 10.3%.

†"Sets" are those dimensions of the testing paradigm which are mutually exclusive in performance, yet interrelate to produce the testing process.

THEORETICAL DIMENSION

The primary assumption underlying the theoretical development of testing was the need to assume that human traits were "inborn" or fixed and that these traits could be observed in physical characteristics and behaviors of human beings. This assumption led to an interest in the possibility of a psychological model which would be based on the need to characterize the distribution and variability of individual differences. Inspired by the work of Galton, Darwin and others, the assumption led to the logical conclusion that quantitative measurement could be devised. Thus, it was conceptualized that highly developed mental traits could be characterized as a set of "intelligent behaviors" or "non-intelligent behaviors"; and that these categories could be applied to "bright" and "dull" individuals respectively. Thus, much of the theoretical development in testing came to be founded in intelligence testing.

Subsequent assumptions have tended only to elaborate on the collectiveness and intricateness of the traits. The first such effort was recorded by Spearman (1927), whose "general factor" ("g" factor) was found to be present in all standardized tests of intelligence, and it was the "g" factor that allowed the measurement of complex mental abilities. Later, Thurstone (1938) developed a multiple-factor analysis, illustrating that Spearman's "g" factor could be defined into a number of primary abilities or tests, such as verbal comprehension, space, reasoning and others. This discovery really provided greater stability for the structure of the "g" factor and this stability was not really questioned until the work of Cattell.

The investigations of Cattell in the 1940's expanded the omnibus "g" factor theory, into a two-factor theory, thereby introducing another determinant of intelligent behaviors. He distinguished between a fluid factor (g_f) which is independent of cultural and educational acquisitions, and a crystallized factor (g_c) which is primarily dependent on cultural knowledge and educational attainment.*

*Cattell, Raymond (1968), p. 58.

Several investigators have been particularly concerned with the structure of the factors underlying intelligent behaviors and have developed models for these factors. Burt (1949), Vernon (1960), and Humphreys (1962) have been particularly interested in hierarchical levels of factors. Guilford (1956) has focused on the "structure" of intelligent behaviors. His work has revealed five kinds of intellect: memory, cognition, convergent thinking, divergent thinking, and evaluation, that can be applied to three types of content: figural, structural and conceptual. As one can conclude, he has been instrumental in constructing a multi-dimensional model that can facilitate any number of combinations including intellect and content.

In summary, the terms "genotype" and "phenotype" can be used to distinguish the theoretical assumptions that have been offered to designate those factors considered determinants of "intelligent behavior". Genotype has been defined to mean those traits that have been mainly inheritable, while phenotype refers to those traits that have modified and molded by environmental influences (Dobzhansky, 1951). Vernon (1965) states that still another factor should be considered when examining the different abilities measured by a particular intelligence test. He contends that each test measures a different set of abilities, and therefore it adds a test-induced factor to the theoretically derived factors of intelligence.

METHODOLOGICAL DIMENSION

The discovery that human behavior could be described in statistical statements was fostered to a great extent by such pioneers as Quelelet and Galton. Galton was influenced by the mathematical formulations of Quelelet, and as a consequence, he devised many measuring instruments in an attempt to describe sensory thresholds of individuals.

Along with the influence of Darwin's writings about selective breeding, Galton formulated two laws that have had a strong impact on many of the methodological procedures used in present day test construction and analysis of test data. Hirsch (1973) describes these two laws as "Law of Ancestral Heritage" and "Law of Regression". Galton's laws imply strong assumptions about normative population

and about the variability of those populations.

Binet extended Galton's concern for sensory measurement to the measurement of higher mental abilities, and therefore, he devised tests involving complex mental tasks. Terman (1916) standardized these tasks and adapted other tasks to produce the Stanford-Binet Intelligence Scale for the American population. The methodological procedures were set in motion, providing the foundation for other kinds of standardized tests, such as achievement, aptitude and diagnostic tests.

In order to understand most of the data collected from the use of intelligence, aptitude or achievement tests, Kerlinger (1964) requires the adherence to certain measurement postulates.*

In theory, the measurement postulate that governs the kind of data that is generated by intelligence, aptitude or achievement tests requires the use of ordinal scaling techniques. In other words, one's test scores can be given rank order values. Ordinal scale data presupposes that there is no absolute zero point that can be designated, nor can one assume that there are equal empirical distances between the scores.

This characteristic has proven to be a source of conflict in present-day testing practice. The conflict arises when most psychological scales, which are essentially ordinal, must be assumed to have equal intervals, a practice which has occurred primarily out of the necessity to use the most robust statistical tools available. However, there is always the recurring problem of being able to adjust the mechanical procedures enough to assure equality of interval, without the expense of loss of interpretability of the data. It has become a significant problem, especially when certain psychological scales have been applied to diverse population groups. It is in the instance of applying interval-scaling techniques to ordinal scale data of very heterogeneous populations, where the serious errors in interpretation may be too costly to be overlooked. At this point, it may be advantageous to look at the theory/practice

*Kerlinger, Fred W., *Foundations of Behavorial Research*, Holt, Rinehart and Winston, Inc., 1964. Chapter 30, p. 420.

conflict to assess those instances when the interpretation of the data is grossly distorted.

The methodological procedures or psychometric applications in testing have been enhanced through the tremendous aid of the computer.* Test format, item selection, complex scoring and the use of multivariate and factor analyses have contributed to the range and depth of the investigation for which testing instruments have been designed. The technological boom may be one of the reasons for the general reluctance to reexamine and possibly revise earlier postulates and subsequent procedures of testing.

FUNCTIONAL DIMENSION

The most important assumption underlying the functional dimension is that tests can predict human behavior to selected criteria. Unfortunately, problems have arisen especially in the field of education because: (a) the selected criteria were imposed equally upon all segments of the population; (b) there was an increasing demand for mass testing, where the role of testing prediction became intricately linked with the role of testing in placement; (c) a conflict arose between the needs and interests of the individual, the groups and the institution; and (d) testing came to be used as a policymaking tool.

The doctrine that the masses have a right to equal educational opportunity has been accepted generally in recent Western history, and that criterion assumes that even the masses are on an equal socio-economic footing. However, this has not always been held to be true. Prior to the mid-17th century, public education in England was meant for a relatively poor group of people, i.e., merchant families, landlords.** However, after this period, the financial burden of schooling shifted educational attainment to the aristocracy. This period of educational history is of particular interest because it set the momentum for modeling the curricula and instructional methods that enabled the upper classes to survive in their own

---

*Hieronymus (1971) discusses the use of technology in today's testing (p. 59).

**Roman (1930).

environment. However, in most of Europe, the masses regained their right to education through the events brought about by the Reformation.

Subsequent historical periods, both in the United States and western Europe, have been responsive to the rights of a person to educational opportunity, but the educational curricula and the methods of teaching traditionally have retained the environmental learning characteristics most suitable for the capabilities of individuals found in the upper social classes.

In spite of the work of John Dewey* where he extolled the virtues of the scientific method for delineating classroom experiences, there has been little overall change in the nature of classroom environmental learning characteristics. Even though it has been found in countless recent studies, especially in the compensatory education programs of the 1960's, that the application of the same criteria to different cultural groups may not be appropriate in view of the different educational experiences of various groups.

The role of testing in prediction and placement has its origins in the Industrial Revolution, about 1830, and in World War I. As American business and industry became concerned about the use of scientific management principles, the educational system became the primary social institution that could effectively sort and place the diversity of talent needed for a growing technological society.

Pressures to produce "objective" testing instruments for military use in both World Wars led to the creation of the U.S. Army's Alfa and Beta Examinations. It was reasoned that a person's capability and potential could be measured and categorized to serve the interests of the individual and the needs of the Army. The Army tests had as their criteria "a high degree of reliability and a moderate degree of validity."** It can be said that the highly specific rationale for mass testing that emerged in the context of the war era continues to dominate the policy and practice of testing even today.

---

*John Dewey, cited in Tesconi and Morris (1972), p. 150.

**Guilford (1946), p. 427.

The ranking of individuals has been in response to a given norm, which happens to be biased towards that of the dominant group. Therefore, minorities can be misplaced because all levels of minority groups usually have not been represented in the establishment of the norm.

When testing is viewed from the perspective of either the individual, the group or the institution, there must be a compromise in the assertion of the needs and interests of each.

Tests have been used primarily to satisfy an individual's needs, such as guidance into an appropriate career opportunity and the identification of strengths and weaknesses in given subject areas. Manning (1968) defines tests which are of concern to the individual as having "guidance", "elective" and "educative" functions.

Tests also have been used within groups to describe comparative relationships of individuals to specific criteria, such as ethnic/racial, social class, sex and age distinctions between groups.* Perhaps this is what Manning meant when he referred to the "societal function" of testing.** Such a use of tests has the effect of denoting societal values and promoting the homogeneity of societal bias.

There are at least three distinct areas in testing which have become prominent for institutions. They are:

- o In social institutions (as in the determination of human performance in education and effectiveness of educational programs).
- o In economic institutions (as in the selection and promotion of personnel in the business sector).
- o In politico-legal institutions (as in the bargaining for visibility by minorities in legislatures and courts).

---

*Please refer to extensive literature summaries provided by the following studies: Eells et al., 1951; Anastasi, 1958 a and b; Lesser, Fifer and Gordon, 1964; Miller and Dreger, 1973 (ethnic, racial and social differences); Terman and Tyler, 1954, Maccoby, 1966 and Kimura, 1973 (sex differences); and Inhelder and Piaget, 1964 and Kamii, 1971 (developmental differences).

**Manning (1968), p. 260.

Tests, in the first two instances, are used primarily to screen and place individuals for specific purposes as well as substantiate necessary policy decisions so that the institutional effectiveness can be insured, and to select and promote individuals in employment. Manning describes these functions as the "prescriptive," "evaluative," and "selective/distributive" functions of testing, respectively.

Attempts have been made to resolve the individual/group/institutional conflict in the legal and legislative arenas, and these mandates have been presented to administrators and researchers for implementation.

Testing has been used as an invaluable asset to the educational policymaker. An effective policymaking capability requires that a test provide the administrator with test scores which display:

(1) Reasonable psychometric stability in the theoretically and operationally determined aspects of reliability and validity.

(2) Relevant interpretive framework to provide equity in treatment of groups, parsimony in allocation of funds and continuity or compatibility with prior research data base.

The problem areas found within the functional dimensions became critical when legal and legislative controls over the use of tests began to dictate the policy interpretations and implementation of the concept of equal educational opportunity.

The court rulings on desegregation in 1954 began the climate for renewed discussion on the subject of equal educational opportunity. The Civil Rights legislation of 1964* intensified that concern by providing, among other directives, a mandate to respond to the use of testing in employee selection.

Kirp (1974) discusses the fact that the constitutional rights

*Huff (1974), p. 246-269.

of students may be infringed by the testing process.* It has been primarily the issue of accessibility to educational facilities that may lead to later job opportunities that had led courts to pass judgments on the use of testing.

Unfortunately, there are too many nuances of interpretations to be found from a reading of the law, and problems arise in attempting to translate the mandates into policy decision. Several federal agencies are in the process of deciphering the meanings of statutes in an attempt to prepare guidelines for testing, primarily in employee selection. Flaugher (1974) highlights some of the problems inherent in recent decisions, and the practical consequences of the legislative mandates and court litigation in terms of policy implementation.

Goodlad (1971) discusses the changing context of equal educational opportunity, pointing out the difference between "quantity and availability." He asks the question: "...how much constitutes a minimum (or later adequate) core and how easy is it to gain access to the system?"** He goes on to explain that it is the last question that "... provides a breeding ground for questions about equal educational opportunity, for example, to what extent and on what basis is access difficult for some individuals and groups?"†

He distinguishes between the terms "educational opportunity" and "equal educational opportunity." It is evident that before court ligation, the historical goal of "equal educational opportunity" was in many respects, realized only with respect to the dominant cultural group. It satisfied their needs, their aspirations and the ultimate aim of drawing the nation under a common educational standard.

Even though there were special classes for those immigrants who needed to learn English, essentially there were no special programs in public schooling for academic tutelage (Brickman and Lehrer, 1972). This is an interesting aspect of that educational era, in view of the number of compensatory educational programs that presently exist as a result of several legislative mandates for poor and minority groups.

---

*Kirp (1974), p. 7-52.

**Goodlad (1971), p. 4

†*Ibid.*, p. 4.

B: THE STUDY OF CULTURAL GROUPS

The existing paradigm of testing was applied initially to studies of group differences and, subsequently, to studies attempting to eliminate group differences.

DEFINING GROUP DIFFERENCES

A preponderance of the literature documenting group testing has come from studies of race comparisons which depict Black and White differences in intelligence and achievement (Porteus and Babcock, 1926; Klineberg, 1935; Shuey, 1958; Miller and Dreger, 1973).

Miller and Dreger describe the historical sequence in which the comparative research on race has occurred:

> "Most of the comparative research on race has been done within a normative framework, with the behavior of whites being the norm for which blacks deviate. Earlier research was directed primarily at attempts to measure and describe these deviations... More recently, differences between the races were interpreted within a social pathology framework... Spread throughout this review is evidence of a turn to another way of looking at differences. We now recognize that in spite of shared values, there are a number of very real cultural differences between blacks and whites, and that these differences cannot be equated with inferiority as they have been in the past."*

Several studies concerning other cultural groups are dispersed throughout the literature. Such studies include Spanish-speaking minorities (Anastasi and deJesus, 1953; Anastasi and Cordova, 1953; Zirkel, 1972), Oriental/Asiatic groups (Porteus, 1939; Lesser, Fifer and Clark, 1964) and Indian-American groups (Klineberg, 1929; Havighurst, Gunther and Pratt, 1946; Anastasi, 1958a).

Most of these studies have revealed statistical results that show the mean average responses of most minority groups to be below the mean average response of the compared dominant group. As a

*Miller and Dreger (1973), p. 1.

result of that statistical benchmark, there has been a continued proliferation of studies identifying specific cultural antecedents that may be responsible for differences between groups, especially as the antecedents are related to performance on mental ability tests. These cultural antecedents have included socio-economic status and mobility, family background and child rearing practices, rural-urban geographic location, segregated-desegregated school environments, and others. A great deal of interest has centered around the study of these cultural influences and their consequences on characteristic patterns of learning ability among these groups, as well as the mean performance levels on school achievement tests.

Lesser, Fifer and Clark (1964), have been concerned with learning patterns among various groups and stressed the fact that certain groups may have a greater advantage to learning if different learning modalities were examined.

Dreger (1973) and L'Abate et.al. (1973) provide literature summaries on most of the significant comparative research on intellectual functioning and educational achievement. Another comprehensive study includes the major Equality of Education Opportunity Survey (EEOS), a national study conducted by Coleman et al. (1966). Other studies have been made by Mosteller and Moynihan (1972) and Jencks et al. (1972). These authors have been instrumental in analyzing massive amounts of data; revealing several determinants of educational inequality. With respect to educational achievement and standardized testing, these data suggest that there is a definite achievement gap between the "dominant" group and "minority" groups and that this gap widens as these groups move through school.

Levine (1972) states that more than 75 percent of pupils, particularly low-income students, are "at least two years below the national average in reading by the time they reach the seventh or eighth grades."* Mayeske (1969)** discusses at least three other types of achievement test score performances for different racial-ethnic,

*Levine in Brickman and Lehrer (1972), p. 42.

**George Mayeske (1969) Technical Paper No. 1 (Office of Program Planning and Evaluation).

regional and socio-economic groups. He states: Verbal ability shows the characteristic decremental learning curve over grades, while Reading Comprehension is almost linear. For all races, mathematics achievement appears to approach a plateau much earlier than other subjects, with the Negro students showing relatively little progress beyond the 9th grade.*

ELIMINATING GROUP DIFFERENCES

Studies attempting to eliminate group differences have their theoretical origins in the numerous efforts to approach the concept of "culture fairness". Each theoretical proposition soon was followed by empirical studies on strategies for reducing cultural bias, and the idea of "fairness" towards groups has been reinforced by political, legal and administrative actions. As a result, certain pedagogical implications have been created.

APPROACHING THE CONCEPT OF "CULTURAL FAIRNESS"

The concept of culture-fair testing has been in existence since 1940, when it was introduced theoretically by Raymond Cattell. However, the reality of the need for a culture-fair testing perspective also was recognized by several previous investigators who supported an adjustment in the interpretations of test scores when applied to various cultural groups. (Klineberg, 1929; Daniel, 1932).

Cattell defined a culture-fair test as having spatial reasoning and numerical test components, emphasizing the non-verbal aspect of mental ability. Previously, it had been held that these components were not primarily influenced by one's cultural background or educational attainment; and, therefore, the test items were considered culturally-fair. Cattell's comparative results suggested that the culture-fair test could be used cross-culturally, as well as within subcultures and social classes.

Other investigators have been influenced by such testing procedures and have capitalized on the use of perceptual forms as the non-verbal component to be used in culture-fair tests. (Raven, 1956; Porteus, 1950).

---

*Ibid., p. 18.

Subsequent research activity in the area of culture-fair testing has been extensive. Several investigators have tried, since Cattell, to contribute to the empirical definition of the concept by offering fine shades of meaning, including the terms "culture-free" and "status-free" or by recommending either the "fair use" or "no use" of tests.

Culture-Free

The term "culture-free" emphasized the process of selecting test items that would have little or no cultural loading. Davis and Elles (1954) attempted to produce a culture-free test by reducing the verbal components of tests through the use of pictures depicting common activities found at all levels of the American society. This test is now non-functional. Other culture-free test constructions were attempted, but most have not enjoyed any success because they did not correlate with other tests (i.e., did not have concurrent validity), nor were they useful in predicting to some commonly used criteria.

At least two explanations have been given for their failure. They are: a) It is difficult or impossible to create a relevant test that is not culturally-loaded so as to satisfy many of the required uses of tests; and b) By failing to change both the cultural loading of the test and the criteria to which the test predict, the concept of "culture-free" was non-functional.

Status-Fair

Jensen (1968) has suggested that the term "status-fair" be used in the place of the term "culture-fair". He believed that the latter term should be used as an anthropological term, one which would invite discussion of truly cross-cultural testing between two or more distinct cultures. However, he feels that present discussions in the United States are centered around social class and ethnic differences within a national culture, and therefore they should be treated in that context.

It can be inferred from his writings that he believes that testing was originally designed from a European, upper-class educational tradition; and that the testing format and content, especially intelligence testing, have always had a built-in class bias, although not necessarily a built-in

cultural bias.

In illustrating this point, the American society was seen as a single national culture with only distinct class and ethnic differences. Therefore, culture-fair was deemed an inappropriate label that should in fact be "status-fair". In his discussion of criteria for establishing "status-fairness", he states that for a test to be judged as being fair, it must be:

- o capable of revealing status differences where such differences are due to genetic factors as well as cultural factors;
- o capable of having predictive validity, whereby the test is not biased in favor of one group over another;
- o capable of revealing lower environmental correlations to test scores;
- o capable of showing resistence to practice gain and minimum transfer across equivalent forms of tests.

Empirical research studies have substantiated the importance of the socio-economic status (SES) variable in testing, especially in areas of intelligence and educational achievement. Unfortunately, however, the differentiation and control of other antecedent factors besides SES, such as ethnicity, sex and demographic characteristics, have characterized most of the studies as emphasizing descriptive methodology instead of experimental methodology.* The former approach has limited much of the potential for generalizability in data involving the study of SES. There has also been considerable concern with the definition of social class indexes between diverse cultural groups.

Culturally-Optimum

Darlington (1971)** uses the concept of "cultural optimality" instead of the concept of "cultural-fairness". He divides the use of the term into two components: a) "a subjective, policy-level question concerning

---

*L'Abate, Oslin, Stone (1973) Comparative Studies of Blacks and Whites.

**Darlington (1971), p. 79.

the optimum balance between criterion performance and cultural factors ..." and b) "a purely empirical question concerning the test's correlation with the culture-modified criterion variable and whether that correlation can be raised." He explains that the concept of culture-fair implies two conflicting assumptions used in test construction and selection: a) the maximizing of test validity and b) the minimizing of the test's discrimination against certain cultural groups.

These two conflicting goals traditionally have been tolerated by constructing tests with a high degree of reliability (discrimination between groups) and a lesser degree of validity. The concept of cultural fairness implies a relative balance in the attainment of the two goals, and that a mechanical advantage of either sets up a critical imbalance and "mutually contradictory definitions" when applied to the concept of "cultural-fairness". Darlington concludes that the choice of the priority of goals is a policy-level decision. Only after that decision has been made can there be a psychometric procedure resulting in the construction of a "culturally-optimum" test.

Research studies concerned with the psychometric constraints on predictive models of testing as they have been applied to the concept of cultural-fairness have been numerous (Cleary, 1968; Linn and Werts, 1971; Thorndike, 1971). The most comprehensive research study on the differing value perspectives of test prediction and their psychometric implications has been completed by Cole (1972). She lists all of the models to date that deal with the definition of culture-fairness or the concept of cultural optimality in the selection of minority group members for employment or college programs.* In seeking a practical application for the Darlington concept, Cole relates it to the problems of employee selection and college admission, but avoids a discussion of its application in early and elementary education.

The weakness in the Darlington concept of "cultural optimality" seems to be that it skirted the theoretical issues of reliability and validity when dealing with different cultural groups.

---

*Cole (1972). List six models: quota, regression, employer, Darlington, Thorndike, equal opportunity models.

Fair Use

Thorndike relates the fairness of a test to its "fair use". As can be noted, he does not restrict himself to the term "cultural" in his definition. He suggests the following:*

> "If one acknowledges that differences in average test performance may exist between population A and B, then a judgment on test-fairness must rest on the inferences that are made from the test rather than on a comparison of mean scores in the two populations."

It can be concluded from Thorndike that "fairness" can be approached on a conceptual basis if we assume that a significant relationship exists within a group, i.e., between a test and its criterion variable. The basis for inferences of whether the test was fair or unfair, therefore, would be in a comparison of the pattern of relationships between the two groups.

In examining the literature on "culture-fair" testing it is within the conceptual analysis of the term "fair use" of tests that a paradigmatic mode of analysis of "culture-fair" emerges. The mode of analysis that presents itself is known as "equivalence", and this term will be discussed more fully in Part II of this paper.

No Use

There have been several calls for a moratorium on testing. The Association of Black Psychologists called for a moratorium on "the repeated abuse and misuse of the so-called conventional psychological tests", as they are "unfair and improperly classify Black children."** The Human Relations Conference of the National Education Association called for a stop to the school testing of minorities.*** Some state legislatures also, such as the California Assembly, have been sensitive to the discriminatory effects of testing.†

---

*Thorndike (1971), p. 63.

**Williams, R. L. "Black Pride, Academic Relevance, and Individual Achievement," *The Counseling Psychologist*, Vol. 2, No. I, 1970, p. 18-22.

***National Education Association, Conference Report on Testing, 1972.

†Mercer, Jane (1974a), p. 138-139.

THE REDUCTION OF CULTURAL BIAS

On the whole, results reported from studies of group differences provided very depressing forecasts for the educational futures of most of the cultural groups that had been investigated. Even though some studies documented in great detail the presence of culture-specific information in tests and the need to eliminate such information (Eells, 1951; Davis and Eells, 1953), the shift toward the elimination of culture-specific information was very gradual. It was not until about two decades later when the theoretical and empirical implications of the study of test bias became an actuality. At that time, the apparent obliquity of test results of the minority group from the test results of the dominant group became less meaningful when the testing instrument itself was brought into question as being biased.

At least three sources of bias have been studied in existing standardized tests. They are predictive, item and test taking biases.

Predictive bias initially was defined to mean that the mathematical model used to explain the behavior of the data predicted more accurately for one group than for another group. Jensen (1968, p. 78) states:

> If a test has different predictive validities for different groups in the population and these differences cannot be attributed to differences in variance on the test or the criterion, it is likely that the test is biased in favor of some groups and not others.

Several investigators have examined the validity coefficients to see if they were the same for various groups. The Educational Testing Service (1966) studied the Preliminary Scholastic Aptitude Test (PSAT) and the Scholastic Aptitude Test (SAT) to find out whether the test scores for Black and White students predicted equally well to the grade point averages in all groups. All groups were in integrated colleges, and the findings suggested that predictor scores for both groups reacted the same way whether placed in the common regression equation or the specific equations of the two groups. However, it was noted that in one of the colleges the grade point average was over-predicted when the common regression equation was used.

Cleary (1966) examined the predictive bias between Black and White

students in integrated colleges. Results revealed that there was no evidence of predictive bias in the tests for the Black students. In her research, Cleary's definition of bias has been stated as:*

> A test is biased for members of a subgroup of the population if, in the prediction of a criterion for which the test is designed, consistent nonzero errors of prediction are made for members of the subgroup.

In other words, if consistent nonzero errors were obtained, the under or over prediction of scores would not be at the disadvantage of either groups involved. This takes for granted the use of a single prediction equation used for both the majority and minority group.

Stanley and Porter (1967) found the predictive validity of the SAT to be about as "correlationally valid" in predominately Black colleges as it is in predominately White colleges. They found the interpretability of the test in Black colleges to be restricted, however, because the distribution of scores displayed a highly skewed curve.

Greene (1974)** reviewed the literature and reported studies which contained contrasting viewpoints about predictive bias; that is, the SAT and ACT (American College Test) were poor predictors of performance among Black students who came from segregated southern high schools and entered integrated colleges (Clark, 1965), and among Black students in predominately White colleges (Bradley, 1967).

Linn and Werts (1971) discuss the problem of predictive bias differently. They state "that the definition of predictive bias requires a comparison of regression equations and is not equivalent to a comparison of validity coefficients." They go on to say that "equal validity coefficients can easily be obtained from quite different regression equations ... therefore given a common regression equation for two or more groups, the within-group validities can be substantially different." These investigators grant that, for this definition to be operational, there must be the assumption that the criterion is free of bias.

---

*Cleary (1968), p. 115.

**Greene (1974), p. 181-182.

Comprehensive reviews of the literature on the bias in prediction models have been given by Cole (1972) and Flaugher (1974). Cole gives a technical summary of what constitutes bias in each of six models based on different psychometric assumptions as well as the valued judgments involved in their selection. Flaugher discusses the definition of bias in each model in non-technical summary and suggests the most probable compromise in the use of one model, given the practical problems of legal interpretations and policy implementation.

Item Bias may be defined as the study of those clusters of items that are particularly easy or difficult for one group when compared to another group. In other words, most studies of item bias are concerned with either the quantitative or qualitative analysis of item difficulty. Quantitative item difficulty refers to the emphasis on rank order analysis where judgments are made about test bias through statistical procedures. Breland (1974) summarizes the operation:

> While these studies are labelled studies of 'item bias', they rarely attempt to analyze sources of deviation for outstanding items. The attempt has been usually to make some inference about the test as a whole by demonstrating the existence or lack of existence of a significant item x group interaction.*

Qualitative item difficulty usually refers to non-technical procedures, whereby clusters of items are judged by their culture-specific informational content. Some empirical verification of culture-specific content is usually cited. One such empirical study was cited by Armstrong (1972), where persons from various ethnic groups were asked to judge those test items that were considered biased toward their group. Even though the kinds of items selected among groups were very different, selections of biased items within each group were similar.

The quantitative aspect of item difficulty analysis can be seen in the Educational Testing Service (1966) study of item bias in PSAT and SAT for Black and White students attending integrated colleges. They found no significant "item x race" or "item x socioeconomic status" inter-

*Breland (1974), p. 4.

actions within groups. It was concluded that items were unbiased. When investigators Stanley and Porter (1967) studied item difficulty levels in SAT for students in predominately White and Black colleges, contrasting results were found. For the Black students, item difficulty levels were inclined much more toward the lower end of the scale, prohibiting a normal distribution of scores. These results revealed that the difficulty level of the SAT was unusually high for this group of students.

Of particular interest in this article was the discussion of item difficulty level and its relation to the predictive validity of the test for the minority group examined. If the difficulty level of the test items is such that subgroup response cannot be subjected to the normal curve distribution model, at least two problems become evident. Either the test items are too difficult for the group or the type of populations that are characteristic in some minority group institutions are different; and alternative, probabilistic models should be investigated to adequately interpret traditional test results. Cleary and Hilton (1968) studied biased test items on PSAT for Black and White students attending integrated colleges. An item on the test was considered biased if the performance on an item by group members differed more than expected between groups on all other items included in the test. Their conclusion was stated as "PSAT items cannot for all practical purposes be considered biased for either race (White or Black) or SES within race".* The phrase "for all practical purposes" seems misleading in that some items were biased according to their definition of item bias.

Angoff and Ford (1973) examined items on the PSAT from Black and White students using correlational analyses to depict item difficulties between groups. They found that some items were unusually difficult for Blacks and went a step further to explain the content of the item. They stated the areas of difficulties to be with "vocabulary and concepts pertaining to unfamiliar places and experiences."

Breland (1974)** studied the cross-cultural stability of test items

---

*Cleary and Hilton (1968), p. 69.

**Breland (1974). Tests were: vocabulary, picture-number, reading, letter-groups, mathematics and mosaic comparisons. The groups were: American Indians, Blacks, Mexican-Americans, Puerto Ricans, other Latin-Americans, Oriental-Americans, White Northeastern, White North Central, White Southern and White Western.

on six different cognitive tests using responses from 10 different groups. This data was received from data already collected by the National Longitudinal Study of the High School Class of 1972, a study of the Educational Testing Service funded by the Office of Education. There was no adjustment made for SES levels, and it was argued that the samples from each cultural group had been randomly selected.

Breland combined a "mechanical and subjective" approach to investigate the instability of test items within each subgroup. By that term he meant the adapted procedure used by Angoff and Ford, where the number of items answered correctly by each group are normalized and to which appropriate delta values are assigned. A cross-plot is constructed and nine cultural groups are compared to the North Central White group. He discusses his correlational analyses to involve the "line of best fit", that is, he defines cross-cultural unstable items as "those with the most aberrancy around the line of best fit for a particular group".*

The test results were not surprising. Vocabulary items were considered most unstable among groups, and this was ascribed to the linguistic varieties of the groups. There were categories in the mathematics test that were relatively easy for the groups, while others were especially difficult. Breland suggested that questions requiring a knowledge about numerical relationships in life situations were less difficult. Certain mathematical problems, such as "determining value of square roots of whole numbers less than ten", were difficult. This conclusion reflected serious problems in the attainment of certain basic mathematical learning in the schools. Most aptly, Breland summarizes the findings, as follows:

> While the cross-cultural stabilities of some item types suggest problems in test construction, instabilities in other item types point to inadequacies in schooling.**

The qualitative aspect of item difficulty analysis can be seen in several literature studies which have been concerned more with the *content* of item bias than with the *technical aspects* of defining item bias. Dis-

---

*Breland (1974), p. 20.

**_Ibid._, p. 51.

cussion found in the investigations of Breland (1974) and Angoff and Ford (1971) embraced both concepts of defining item bias: "mechanical" (statistical analyses) and "subjective" (face-valid content) judgments. Other investigations in item bias have included the reasons for response choice, culture-specific informational content in the test item and patterns of abilities reflected in the choice of item response.

Brigham (1932) laid the foundation to the study or correct/wrong responses among distractors by suggesting that incorrect responses were deliberately chosen instead of being selected at random. In other words, he provided an alternative method for studying biased item response. Theoretically, Brighams' work with the College Entrance Examinations Board provides a framework in which to understand item response. This framework could have an invaluable impact in understanding item choice in various cultural groups where total "correct" responses are lower than the dominant group. Brigham states:

> "It is possible to show that items which apparently have hundreds of possible answers, instead of five, show certain characteristic distributions of answers indicating concentration of errors."*

He goes on to point out:

> ". . . that the ultimate facts with which we are dealing are answers to questions. It is not necessary that these answers be scored or have values attached to them by some tester -- the answers may be studied in their own right.
> . . . the detailed study of answers to test items provides a completely sound and systematic approach to the study of errors and confusions in thinking."**

Later on in his text, he suggests "that we are nearer the truth in conceiving of 'intelligence tests' as measuring the degree of participation in the group mind . . ." and that "symbolic manipulations are not

---

*Brigham (1932), p. 43.

**_Ibid._, p. 45.

random phenomena but subject to social control."*

Several investigators have used error item analyses as indices to explain influences and classify cultural differences between groups (Eells et al. 1951; Lawrence, 1957). The study of Eells and his colleagues was primarily concerned with "intercultural differences among white groups" from the point of view of content, these authors were interested in examining those items which revealed:

- o unusually large status differences
- o unusually small status differences
- o sets of items showing contrasting amounts of status difference although similar with respect to form of symbolism. (letters, pictures, numbers and type of question)
- o significant differences between two low status groups (old American and Ethnic).

With reference specifically to content, vocabulary items were stressed to be most important in dividing the cultural groups. In summary, these authors state:

> Practically all of the items which show unusually small differences either are non-verbal in symbolism or are expressed in relatively simple everyday vocabulary and deal with objects or concepts which are probably equally familiar, or equally unfamilia: to pupils of both status levels.

Another finding suggested that "there were a large substantial number of items showing large status differences for which no reasonable explanation was noted." It was advised in this instance that caution should be taken "in accepting the idea that all status differences on test items can be readily accounted for in terms of the cultural bias of their content."**

From a different perspective, Roberts (1970) summarizes an evaluation of linguistic item biases found in four tests that are used frequently to measure language development and abilities in young children: The Peabody

---

*Ibid., p. 208.

**Eells (1951), p. 357.

Vocabulary Test, Wechsler Pre-School Primary School Intelligence Test, Metropolitan Readiness Test and Illinois Test of Psycholinguistic Abilities. She points out that "substantive bias in standardized tests can be found in culture-specific vocabulary items, culture-specific pictures, culture-specific information questions and even dialect-specific linguistic questions.".*

Wolfram (1974) cites examples of cultural bias in diagnostic tests for articulatory development, auditory discrimination, grammatical development and vocabulary acquisition. From a sociolinguistic point of view, his concept of task bias embraces comprehension of instructions and interpretations of an appropriate response set. It also includes specific linguistic item bias found in the phonological and lexical differences between dialect responses and test commands given in Standard English.

Other studies have revealed linguistic item bias in standardized tests used in grade school in reading (Meir, 1973) and other subject areas (Cicourel et al., 1970).

Lesser, Fifer and Clark (1964) attempted to reduce cultural content in their study of six and seven year old children from two social classes (middle and lower) and four cultural backgrounds (Chinese, Jews, Blacks and Puerto Ricans).

Their "culture-fair" materials was described to "presuppose only experiences that are common and familiar within all of the different social class and ethnic groups in an urban area." One finding in this study was that after the item as stimulus was controlled for cultural differences between groups, patterns of abilities among groups remained different for each ethnic group. In addition, the authors state that "once the pattern specific to the ethnic group emerges, social class variations within the ethnic group do not alter this basic organization".**

The results of the studies appear to be inconclusive in their attempts to detect and/or remove cultural bias found in existing tests. This has been so because the primary objective of the studies has been to eliminate differences in performance between groups, not considering that these differences may not be manipulatable through technical analysis or change of

---

*Roberts., p. IV-13.

**Lesser, et al., p. 567.

content. The overwhelming problem seems to be in the interpretation of these salient differences, so as to make valid inferential statements about the test results among groups. To date, this question has not been fully explored in the existing literature studies.

Test-taking Bias may be defined as a mismatch between the normative expectations of the test designer/examiner and the personality factors and learned skills of the test taker. The use of standardized testing to measure psychological and educational behaviors is accompanied by a set of standardized test-taking behaviors. In other words, standardization is not only controlled through externally valued criteria.* Briefly, such criteria may include the ability to follow instructions; the ability to work persistently and/or speedily through a series of tasks; and the ability to manipulate numerical, geometrical and linguistic relationships. These criteria provide the necessary framework from which the test constructor/examiner must design standardized test-taking norms which may be at complete odds with the personality factors and the acquired skills of the intended test-taker.

Several labels have been used in the literature to describe these test-taking behaviors. Jensen (1968) provides the following summary: "'motivation', 'test anxiety', 'test sophistication' and other test-taking attitudes, 'personal tempo', 'clerical skills' and 'susceptibility to distraction'".**

A fairly large body of research has been done on examiner bias in testing especially when the race of the examiner is different from that of the person being examined. (Rosenthal, 1966; Sattler, 1970; Epps, 1974). Some attention has been given to "subject bias" in psychological research in general (Lester, 1969) and in the standardized testing situation in particular (MacKay, 1970; Roberts, 1970; Wolfram, 1974).

Rychlak (1973) offers a theoretical learning framework that can be very useful in explaining some of the traditional assumptions of testing. He makes the following case:

---

*These criteria have been discussed on several occasions; Jensen (1968), Brickman and Lehrer(1972), and Jencks (1972).

**Jensen (1968), p. 70.

> It is pure fiction to assume, as many E's [experimenters] do, that S's [subjects] conceptions of the experimental purpose (i.e., design) are 'chance' variations to be cancelled out by another S's conceptions. A major aspect of the learning going on in all human studies has to do with the informal study being conducted by S as to 'what is this all about?' This is literally a controlled dimension amounting to a kind of social role (or rule) which enters into the differential variance accounting for significance in the eventual statistical tests.*

Much of the test-taking bias can be explained from a socio-linguistic point of view given the fact that one's language use and styles provide sufficient familiarity with the type of tasks and the pattern of response required in a particular testing situation. Many literature studies have documented the "social control" of language use and style in many standardized testing situations.

Roberts (1970) states that the "verbal style required by the test can be culture specific". She gives the example that the cultural norms for verbal interchange may be very different from the norms of the test-takers own "speech community".

MacKay (1970) believes the manipulation of a subject's test-taking behavior is based on at least two assumptions. First, the need of the test designer to envision a model in which all of the subjects' actions are predictable. Second, that the subjects' actions can be controlled through testing format and procedures. Summarily, MacKay points out that testing theories are based on the assumption that the administration of the test will take place "in a non-contextual social setting with a non-contextual cognitive orientation".

L'Abate, et al. (1973) summarizes the non-intellectual factors that seem to influence achievement testing outcomes as: self-concept, motivation, level of aspiration, attitudes, etc. Even though, these factors are deserving of research study in their own right, barring the inadequacies of theory and methodological procedures, these variables are said to be present during the complex testing process, and their measurement must be included because they are considered additional sources of variation.

---

*Rychlak (1973), p. 3.

Brigham (1932) seems to be discussing test taking bias as a group phenomena when he used the terms "intrinsic causes of group factors". He believed that "it was possible to show that group factors may either be suppressed or generated by experimental conditions of testing, such as timing. . . ." He concludes by stating:

> "There may be other irrelevant testing conditions set which tend to alter the results one way or another. The study of these conditions by experimental variation and control is a most important problem and one which should take precedence over the mathematical systems of interpretation which have now gone far beyond the test data."*

THE POLITICAL, LEGAL AND ADMINISTRATIVE ACTIONS

Efforts to achieve cultural fairness were reinforced by political, legal and administrative actions.

Political Actions. The concept of cultural fairness in testing was extended from a scientific/academic debate to the political forum as a result of numerous writings suggesting national inquiry into the use of tests (Hoffmann, 1962, and Black, 1963) and as a result of specific federal legislative mandates (EOA Act of 1965 and ESEA of 1965) and state legislative mandates.**

The "fair use" of standardized tests, in relation to various groups, has been unalterably associated with the two concepts of "equal educational opportunity"*** and "educational accountability".† The curious

---

*Brigham (1932), p. 44.

**Please refer to Clasby, Webster and White (1973) for extensive summary of state legislative mandates authorizing the use of tests to assess educational programs.

***Goodlad, J. (1971) distinguishes between the terms "educational opportunity" and "equal educational opportunity" through a historical, social and economic context. He explains these terms in changing contexts of educational history.

†"Educational accountability" can be defined as the demand on various funding sources to press educational systems for reliable information on student learning to justify the allocation of resources and educational expenses. [Please refer to Tyler (1973) and Webster (1973) for specific rationales for the renewed interest in the need for present-day educational accountability.]

connection between the use of standardized testing and those concepts is highlighted to gradual prominence when one peruses the goals of nation-wide and state-wide testing programs.

There has been a heavy "federal initiative" in sponsoring compensatory educational programs for poor and minority groups. The literature abounds with programs, plans and experiments to give many such children an equal chance in the educational system. However, compensatory education program objectives, as measured by standardized tests, have only revealed short gains that have not been sustained for long periods of time.* The use of standardized tests to assess educational program outcomes has had mixed reviews in the literature. It became increasingly clear that test scores could not be translated easily into program objectives, for two reasons: First, many of the standardized tests used were not originally designed or intended as evaluative tools; and secondly, the utility of aggregated test scores as sole indicator of the effectiveness of the programs left much to be desired.

Millions of dollars are being spent in federally funded research, development and evaluation projects that concern the quality of education of young, minority children. The high concentration of minorities in programs for the poor has highlighted issues of testing in public debate. Such programs as Head Start in early childhood education and Title I programs in elementary and secondary education have been created through the mandates of such legislation as the Economic Opportunity Act (EOA of 1964) and Elementary Secondary Education Act (ESEA of 1965).

Head Start has a current budget of $400 million, and the Nixon Administration suggested that there be a 10 percent increase in the coming fiscal year. As a result of ESEA, Title 1, nearly $1 billion have been allotted to schools with concentrations of children from homes in poverty, and the Act requires local districts to evaluate the effectiveness of the educational programs that emerge. With the caveat of evaluation of the educational programs added as an obligation of the successful execution of federally-funded projects, the need for general guidelines for federal

*Please refer to the extensive studies involving national evaluations of compensatory educational programs (Cicarelli, 1969; Coleman et al., 1966; Follow-Through Evaluation, 1973).

policy in the screening of the selection and use of tests seems to be of paramount importance.

In many instances, there has been a concerted effort not to equate standardized testing with the total design of the evaluation. In spite of the fact that the state-of-the-art of the testing of young children generally is in a fluid state, there has been a tendency to rely too heavily on the results of standardized testing, especially when dealing with diverse cultural groups.

A noteworthy statement was made by Campbell and Erlebacker (1970) as they discuss previous evaluations of compensatory educational projects. They state that "commitment to reality testing (referring to true experiments) on ameliorative programs should involve acceptance of the fact that some programs will turn out to be ineffective." They go on to state that "when such outcomes are encountered, the political system should seek alternative approaches to solving the same problem, rather than abandon all remedial efforts."*

There has been an increasing growth in state legislation authorizing state-wide testing programs of schools and school systems.** The response of the States to the primary "federal initiative" has been to introduce several versions of accountability legislation.*** Webster (1973) records and studies approximately 54 pieces of legislation. She states that 34 were dated in 1971 or 1972, and 12 in 1969 or 1970. She concludes that "over 80% of the legislation was introduced in the past four years."

The problem with this influx of state-wide testing programs has been the undue reliance on test-related information to support policy decisions. This has been especially noticable with the granting of financial rewards. Dyer and Rosenthal (1973) break down this problem into four salient questions: †

---

*Campbell and Erlebacker (1970), p. 203.

**Maureen Webster and Naomi White (1973) discuss "minimal skills", state-wide educational assessment programs and changing context of educational policy.

***Webster (1973), p. 65.

†Dyer and Rosenthal (1973), p. 122.

- Does one use the funds to reward the districts that show up high on the indicators?
- Does one withhold the funds to punish the districts that show up low on the indicators?
- Does one use the funds to help upgrade the districts that show up low on the indicators and thereby withhold funds from those that show up high?
- Or can one find a way to allocate the funds so that all districts will have an incentive for constantly improving the quality of their schools?

The complexity of this problem reveals the varied emphasis given to the role of standardized testing in each state and the necessary linkage between federal administrative policy on evaluation and the general use of tests to evaluate educational objectives at the state level. Since the constitutional authority for education lies in the domain of each State, it is the responsibility of each State to resolve the question of what criteria it will use to judge equitable education performance, and it is the responsibility of the State to make sure that the chosen criteria do not systematically discriminate against certain groups more than others.

Necessarily, the federal policy-maker will be concerned with policy options involving testing alternatives while the state policy-maker must reckon with policy analysis and the implementation of testing objectives. Two trends make a collaborative venture important to both federal and state administrative agencies. First, more than 75% of current state assessment programs rely totally or partially on federal funds. This may be modified in part by revenue-sharing funding proposals in education. Secondly, the necessary distinctions between "federal educational policy" and "national educational policy" Webster (1973) suggests:

> The phenomenon of national coalitions has reached a point where it is possible to distinguish, at least conceptually, between federal educational policy which guides the activity of the federal government and national educational policy positions which represent a wide array of concerns of interest groups and decision-makers.*

*Webster (1973), p. 53.

The latter group will represent an interplay between both federal and state assessment activities. It may be within this realm that the use of testing as tool will be put in its proper perspective.

Legal Actions. Existing tests have been documented to have systematically discriminated against minority and poor children so that they appear to perform poorly on a variety of tests under various circumstances. This documentation can be cited in various class-action suits and court decisions (Mercer, 1974a, and Williams, 1971).

Robert L. Williams (1971) provides examples of some of the racially discriminatory effects of testing. They are summarized below:

- . . . case of *Diana et al. vs. California State Board of Education* led to a decision in favor of a Mexican-American child whose intelligence had been woefully underestimated by the Binet . . .
- . . . case of *Hobson vs. Hansen* in Washington, D.C., set an early precedent in the decision ordering the track system to be abolished since unfair ability tests were used in sorting the children into tracks. . .
- . . . the case of *Stewart et al. vs. Phillips et al.*, charges that children are being placed in special classes irrationally and unfairly . .
- . . . case of *Armstead et al. vs. Mississippi Municipal Separate School District et al.* involved the use of the GRE for employment and retention of Black and White teachers.

David Kirp (1974) provides a discussion of the sorting of individuals by educational institutions that has led to "judicial inquiry". He gives particular interest to "exclusion", "ability grouping" and "assignment to special education".

Volumes of filed suits of discriminatory hiring practices because of testing can be found in the archives of the Equal Educational Opportunity Commission in Washington, D.C. To date, a summary of this literature has not been attempted.

Administrative Actions. Administrators at the federal and state levels have sought to satisfy the above requirements of political and legal jurisdictions, but in doing so they have found themselves in the dilemma of trying to meet the demands of minorities without the necessary theory and data for effective and equitable program implementation.

One of the demands of minorities has been for a more accurate labelling and placement of minorities. Also to be considered are the emotional effects on these individuals of such placement, and the denial of future educational and job opportunities that may arise. Again, Robert L. Williams (1971) cites examples, as summarized below:

- A document from a group of Black psychologists reviewed by the Unified School District of San Francisco illustrated that although Black children comprised only 27.8 percent of the total student population in San Francisco Unified Schools, they comprised 47.4 percent of all students in educationally handicapped classes and 53.3 percent of all students in educable mentally handicapped classes.
- In another instance in St. Louis, during the academic year of 1968-1969, Blacks comprised approximately 63.6 percent of the school population, whereas Whites comprised 36.4 percent. Of 4,020 children in Special Education, 2,975 (76%) were Black; only 1,045 (24%) were White.

Jane Mercer (1974b) supports the view that a greater chance of "mislabelling and erroneous placement" increases as one's milieu at home differs from the cultural milieu of the school. She estimated that "at least 70 percent of the children in classes for the educable mentally retarded in two southern California school districts were mislabelled as mentally retarded."*

Testing of Spanish-surnamed children has intensified the debate over the discriminatory effects of testing. Zirkel (1972) describes existing literature to reveal that there are linguistic, cultural and psychological difficulties for Spanish-speaking children on standardized tests of

*Mercer (1974b), p. 6.

ability and achievement. The language/cultural references made in the test content and the frustration of translating the subtleties of the English language into appropriate Spanish adaptations are the primary variables that have been found to be discriminatory against many Spanish speaking children performing on well-known standardized tests.

Looking back on the early 1960's, many critics have discussed the various problems involving the lack of necessary instruments, strategies or data needed to implement programs relevant to various cultural groups. There was a lack of basic knowledge about the lifestyles and the educational problems of the minority groups (Berke and Kirst, 1972). There was a deficiency in the interpretative framework (existing monocultural paradigm of testing) which could not support the conclusions drawn about these various groups. Unfortunately, this framework was dependent heavily on the results of standardized testing, and often time new programs would show unfavorable results (Fein and Clark-Stewart, 1972). There was also a need to make immediate decisions about strategies for the implementation of program goals before a format or "social experimentation" had been empirically verified (Timpane, 1970). Naturally, such a structured experimentation would have provided, at least in part, the empirical base for needed policy decisions.

Because of the aforementioned reasons, and others, testing was considered a "dependable" administrative tool which, under the existing shortage of information, could provide reliable and valid data about the performance of various cultural groups. In reality, the administrative level to which testing is most helpful is debatable. Nevertheless, the effects of the interpretability of aggregate test scores must be weighed in a broad perspective. Traditionally, standardized testing has not provided this kind of perspective and various cultural groups have been considered at a distinct disadvantage when this kind of testing has been used. More often than not, the interpretability of the test results continues to be considerably influenced by the established cultural norm.

Klitgaard (1974) provides a set of alternatives in the use and interpretation of test measures and statistics that may be of interest to the decision-maker. These alternatives were to demonstrate the theoretical feasibility of interpreting achievement data beyond test score averages

to the examination of the "distribution of scores". The parent study involved the Education Voucher Demonstration Project which supported a variety of objectives including "increased parental influence and satisfaction with schools", "more diversity of educational programs" and "ultimately better education".

It may be said that culture-fair testing as a strategy has found itself in a reactionary position; that is, it has attempted to change the existing testing format, content and psychometric operations. At best, these attempts have been inaugurated slowly, many times with discouraging results. Has the concept of culture-fair testing been doomed a failure? The answer to this question is unclear as one reviews the literature. However, it may be said that there are certain pedagogical implications.

THE PEDAGOGICAL IMPLICATIONS

The application of efforts to reduce cultural bias has acted as an impetus for certain pedagogical implications in the form of several educational testing formats and procedures. These implications have had the effect of minimizing the use of standardized testing, while at the same time embracing the goals of culture-fair testing through a de-emphasis on the use of norms for the dominant group as the standardized reference. At least two of these formats and procedures will be discussed in this section. They are: criterion-referenced testing, with particular reference to the National Assessment of Educational Progress (NAEP); and the System of Multi-Cultural Pluralist Assessment (SOMPA).

Criterion-Referenced Testing is by no means new in the field of testing,* but certainly it has become today a viable alternative to the much-debated use of traditional testing by norm-referenced standardized tests. This approach has been seen as a vehicle to reinforce "culture-fair" testing goals in that group performances are not compared to a "standardization group".** Instead, group performances are assessed

*Airasian and Madaus (1974), p. 78, cites E. L. Thorndike as discussing the difference between norm- and criterion-referenced tests in 1913.

**Refer to Airasian and Madaus (1974) for an exposition of trends leading to the use of criterion-referenced measures (p. 76-77).

through the attainment of "criterion skills"; therefore, they are not dependent upon the performance of previous groups for interpretation.

One example of national import that has supported the use of criterion-referenced exercises is the National Assessment of Educational Progress (NAEP) (1969-1975).* The test instruments created in this program have not been designed as standardized tests in the traditional sense. For instance, tests are not used only to generate scores, but are considered exercises that are reported in population group percentages. Finley (1974) distinguishes between the National Assessment program and traditional standardized testing programs. Briefly, these differences have been described in the following ways:

- o exercises of group versus average performance of students,
- o time is extended to 6 to 8 hours rather than 30 to 70 minutes so speed is not necessarily a factor,
- o response set includes a wide variety of stimuli instead of only the pencil and paper variety,
- o exercises are administered to small groups and interviews, not just total classes,
- o exercises are prepared for high and low students, not just the average individual,
- o total scores reflect the number of students who get the correct responses instead of the number of correct responses by a particular student, and
- o results are reported by the various exercises used instead of in relation to a "standardization group".**

System of Multi-Cultural Pluralistic Assessment (SOMPA). Another challenging educational testing program is being designed by Jane Mercer and her colleagues at the University of California (Riverside), called the System of Multi-Cultural Pluralistic Assessment. Even though this comprehensive assessment procedure includes "measures of adaptive behavior and social role performance in non-academic settings", and "a careful screening for physical disabilities", special interest is given here to

---

*Refer to papers by Ralph Tyler, Carmen Finley and George Johnson in "Part Five: Assessing The Educational Achievement of Institutions", Tyler and Wolf, eds. (1974), *Crucial Issues of Testing*, pp. 91-104.

**Finley (1973), pp. 97-98.

the emphasis on pluralistic norms used in this assessment system. Mercer (1974b) reported from initial data that:

> We have tested 2,100 California public school children five through eleven years of age -- 700 Black, 700 Chicano/Latin, and 700 Anglo-American ... Altogether, we tested children in ninety-one different school districts and over 150 different schools ... We factor-analyzed the forty questions asked the mother about the family background and identified nine characteristics of the child's socialization milieu which are relatively independent variables. We found that five of these factors could account for 27% of the variance in Verbal IQ, 13% of the variance in Performance IQ, and 24% of the variance in Full Scale IQ.*

Even though the traditional standardized tests are being used (1973 revision of the WISC and a diagnostic instrument, the Bender-Gestalt), the interpretability of the testing scores should be greatly enhanced through the use of the pluralistic norms and the use of background data provided by the assessment of the socio-cultural environments of the various groups being studied.

---

*Mercer (1974b), p. 14.

PART II: THE PROPOSITION

INTRODUCTION

It is proposed that a cross-cultural comparative paradigm be developed for use in educational testing for minority groups.

It has been shown that the development of the monocultural testing paradigm established an inherent "separateness" of the dominant versus the minority groups in the American society. Considerable data, theorizing and rhetoric have reinforced this conclusion, and problems in interpreting the educational performance of these separate cultures by the use of standardized tests have not been resolved through the monocultural paradigm.

It may be appropriate to consider another paradigm when dealing with diverse cultural groups. Kuhn (1962) suggests that new paradigms are formulated and acknowledged when conflicting solutions to pressing problems develop; that is, new strategies appear to answer more questions than did the previous strategy. Researchers frequently return to original postulates and hypotheses and reexamine them when they cannot be justified by empirical data already collected, and such an approach may be warranted in the approach to "culturally-fair" testing. It may be appropriate, therefore, to consider the American experience through a cross-cultural paradigm when dealing with the subject of testing.

This section recommends the adoption of a cross-cultural comparative paradigm for testing as a means of enabling policy-makers to deal fairly with the "reality" of cultural separation or homogeneity among certain groups. Traditionally, the term cross-cultural has meant the study of distinct cultures from different countries, nations or geographic localities. In this paper, the term "cross-cultural" will refer to the study of different cultural groups within the national cultural milieu of America.

A: DEVELOPING A CROSS-CULTURAL COMPARATIVE PARADIGM

The idea of a cross-cultural comparative paradigm is considered here as having both theoretical and measurement premises, and this section concludes with suggestions on the procedures for creating the paradigm.

THE THEORETICAL ORIENTATION

A cross-cultural comparative paradigm must have as its theoretical base the assumption that cultural groups can be used as variables. Theoretical statements then can be made about these groups, and such theoretical statements must conform to the general scientific goals of being accurate, parsimonious, general and causal.

Support for this position may be found in the work of Przeworski and Teune (1970), when they discuss the logic of comparative research. They assert that theoretical statements can be made about social groups or "systems" or "system-level variables", if those variables are substituted for the "proper names" of those systems.

When discussing social systems they are referring to nations and countries, but they suggest that the principles are applicable to research designs or mathematical models dealing with social science phenomena (e.g., cultural groups).

These authors enumerate at least two problems that are encountered when examining the behavior of variables within systems and at the system level. They are:

> (1) "distinguish between 'spurious' and 'true' correlations when relationships are observed at different levels" (within or between systems).
>
> (2) "distinguish the effects of the variables observable only at the level of systems (diffusion patterns and settings) from variables aggregated from within-system observations (contexts)."*

*Przeworski and Teune (1970), p. 72.

In the first instance, there is some discussion about the gains and losses in the generality of theoretical interpretations of relationships between variables when certain statistical methods are used. Several criteria are submitted to explain the conditions of "spuriousness" when within-system regression equations are the same or different from total regression equations. Since there is always a compromise between theory and measurement, the assumption is made in this discussion by the authors "that within-system relationships are linear or, in other words, that there are no interaction effects *at the individual level*." It will be seen later in the measurement context of this paper that this assumption is usually unfounded.

In the second instance, to understand this problem it must be realized that the authors are interested primarily in those systemic factors "that may potentially influence or be influenced by within-system behaviors, not with properties of systems as potential variables in system or group-level analyses."* They have summarized the factors of interest to include: "diffusion patterns",** "settings"*** and "contexts."

The discussion of "contexts" as systemic factors has particular relevance in this part of the paper because of its emphasis on aggregate individual data and the measurement of that data. Przeworski and Teune define systemic factors as contexts, noting that "when the characteristics of individuals -- whether predispositional, behavioral, or relational -- are aggregated, the social system of which they are members acquires a parameter."† Two contextual variables are

---

Ibid., p. 51.

**Diffusion patterns describe those relationships that may result from "historical learning" sometimes referred to as Galton's problem. (Refer to Przeworski and Teune (1970, p. 51-53), also to the reference cited, Naroll, R., "Galton's Problem: The Logic of Cross-Cultural Analysis," *Social Research*, 32, 1965.)

***Settings are described as "neither diffusional patterns nor aggregates of observations," but "... characteristics (historical, institutional, external, behavior and physical) to which all individuals within a system are, at least potentially, exposed." (Przeworski and Teune, 1970, p. 53-54).

†Przeworski and Teune (1970), p. 56.

distinguishable: those made up of "aggregates of relational properties" and "aggregates of individual properties." It is the former, aggregates of relational properties, that will be of major concern in the following section on measurement, both from the position of equivalence and its application to the principles of psychometric theory.

THE MEASUREMENT CONTEXT

The measurement context of a cross-cultural comparative paradigm is founded on the premise that the pattern of relationships between test responses can be manipulated so as to make accurate inferential statements about the psychological traits being sought. Heretofore, it has been customary to manipulate only the test scores to that end.

In order to deal with patterns of relationships, one must confront both the concept of equivalence and the concept of construct validity.

THE CONCEPT OF EQUIVALENCE

Equivalence is the inference drawn when it is found that there are parallel factor distribution patterns between groups, between sets of test items or between subsets of test items.

The rationale for this position is supported by Przeworski and Teune when they state that "the criterion for inferring the equivalence of measurement instruments can be found in the structure of the indicators" and that the "basic datum" in the comparison between systems is found in the "within-system relationships."

It has been demonstrated in Part I of this paper that attempts to produce a culture-fair measurement instrument revealed considerable system interferences. In other words, efforts were not made to determine whether parallel statements could be made about factor distribution patterns within systems before proceeding to a cross-system analysis. As Przeworski and Teune point out, the comparison of relationships within systems revealing the behavior of items within that system is more indicative of system interference than is the aggregate of system scores.

Another assumption that must be highlighted in the concept of equivalence of these authors is that direct measurements of phenomena are *accurate*. However, this assumption is questioned in greater detail in the following section on construct validity, where it is shown that indirect measurements must be used; and even then, it can only be said that if indirect measurements can be found to be accurate, then it is feasible to proceed toward statements of equivalence.

THE CONCEPT OF CONSTRUCT VALIDITY

Before the notion of equivalence can be suggested as a viable alternative in the measurement of cultural groups, there must be some attempt to relate this term to the principles of traditional test theory and operations. This association between equivalence and psychometric testing is imperative since both seek to examine "patterns of differences" in group responses.

This section of the paper will be interested primarily in the logic of test construction and test operations, and Loevinger (1967) provides an explanation in this regard when discussing objective tests and their role as instruments of psychological theory. She extends the meaning of construct validity to include some of the crucial criteria needed to provide a psychometric foundation for the operationalization of equivalence.

Loevinger examines the roles of the two primary concepts of psychometric test theory, reliability and validity, concentrating on the latter concept as the one that can impart the greatest contributions to psychometric and psychological theory development. She criticizes the classical definition of validity of being "... too vague, too remote from actual measuring operations, to be useful..." Yet, she concedes that it is this definition, the extent to which a test measures what it is supposed to measure, that is most used in the psychometric tradition. In brief, she contends that "... predictive, concurrent and content validities are essentially *ad hoc*," and that "...construct validity is the whole of validity from a scientific point of view."

Loevinger believes that construct validity has two meanings:

- o That the test measures something systematically.
- o That there should be evidence of the particular interpretation of what it measures.

In other words, one can describe these meanings to be defined in the first instance as the "intrinsic validity of the test" and, in the second instance, the "validity of interpretation." She considers the first meaning to include "the degree of internal structure of the items and the magnitude of external correlations" (psychometric criteria). The latter meaning includes "the nature of the structure, content of items, and the nature of the external relations" (psychological criteria).

Loevinger conceives of construct validity as made up of three components*:

- o The "substantive component of validity is the extent to which the content of the items included in (and excluded from?) the test can be accounted for in terms of the trait believed to be measured and the context of measurement. Context includes psychological theory and, in particular, the psychology of objective test behavior."
- o The "structural component of validity refers to the extent to which structural relations between test items parallel the structural relations of other manifestations of the trait being measured."
- o The external component of validity "... concerns correlation with total score. The method of constructing a total score from the item pattern necessarily implies a commitment about the structure of the items, and thus about the

*Loevinger (1967), p. 97.

> structure of the trait measured. That is, in a cumulative test, where the total score is the number of scored plus, an additive model is implied..."

Loevinger's explication of the components of construct validity has been closely allied with the stages of test construction, and this section applies her concepts of structural and external validity to the notion of equivalence.

In the discussion about the structural component of construct validity, it was noted that traditional psychometric theory for the most part, has been preoccupied with the efficacy of total scores almost to the total exclusion of analysis of individual item response clusters. Therefore, it can be readily concluded that the assumptions concerning the structural relations of responses are not routinely validated between groups. The use of structurally equivalent measurements would seem to be indicated as important at this stage of test construction and development.

In the discussion concerning the external component of construct validity, the "non-test criterion," against which the test must inevitably be judged, seems to encourage the formulation of hypotheses about relationships between groups. Predictions about these relationships may have to be adjusted in view of the structural relations found within and between group responses. Even though Loevinger describes the use of factor analysis in this area, serious judgmental decisions may have to be made about the empirical criteria to which these tests predict. The concept of equivalence reveals the need to examine the ways in which empirical criteria are established to maximize test predictive validity, and vice versa.

PROCEDURE

The procedure for creating a paradigm for use in cross-cultural comparative analyses is demonstrated in the following six steps:

FIRST: Select populations to be sampled.

As Przeworski and Teune point out the populations to be sampled in comparative studies should be taken from "natural" groupings, based on societies, economics, politics or culture.* It has been demonstrated that "natural" groupings in the American society would be the minority groups which are the concern of this paper. However, these authors point out that one must be assured that the characteristics of persons selected are sufficiently random to make an adequate sample.

It should be noted that traditional testing techniques have been very successful in identifying which groups should be sampled. Przeworski and Teune suggest that several studies of group characteristics are avaliable.** These studies have delineated some characteristics that may be appropriate for further examination in the context of specific cultural groups.

SECOND: Select behavioral constructs to be sampled within each cultural group.

Behavioral constructs are those stimuli or variables which are used in the construction of test items and are thus operationally defined by the test designer.

Sears (1961) was concerned about the problem of conceptual equivalence and the need to find transcultural variables. Even though he was writing in a broad cross-cultural sense (cultures from different countries), he makes relevant statements that could apply in this situation. Sears points out the necessity for what he calls "transcultural" variables and identifies the criterion essential for their selection, as follows: "They must be measurable in whatever culture is chosen, whether the culture be a unit of the sample population or a source of systematic variation of an interaction variable."† Recognizing that the criteria to be

---

*Przeworski and Teune, *op. cit.*, p. 57.

**Lazersfeld and Rosenberg, *The Language of Social Research*, Free Press, Glencoe, Ill. 1955 (Section IV). Also cited in the same reference is R. Cattell's work "Types of Group Characteristics."

†Sears (1961), p. 446.

used in developing "conceptual equivalence" are not clearly understood, he suggests that in defining these criteria the problems will probably be no different at the cross-cultural level than at the "inter-individual" level.*

Suitable criteria that may be used in the sampling of behavioral constructs can be derived by the factor analytic method. This method has been characterized as a tool for the study of construct validity and more specifically, for the testing of hypotheses about relationships among known variables.

Kerlinger (1964) discusses this method as it has been used in psychological and educational research efforts. He makes the case that little is known about the construct of achievement, and that because in many respects standardized achievement tests are "factorially complex", users should be particularly alert to question their construct validity.**

Loevinger (1967) discusses a problem that occurs at this stage of test construction:

> ". . . the more one objectifies the nature of the universe from which the sample of items is to be drawn, the less likely is the universe to represent exactly the trait which the investigator wishes to measure. Moreover, for any given trait name, two investigators would not necessarily specify the same objective domain for which to draw a sample, nor the same method of sampling."†

THIRD: Establish criteria for the method of sampling behavioral constructs.

Literature is available which is devoted to the study of the methods of sampling test responses. Such literature is devoted primarily to the various approaches used to get the most reliable and valid responses from children.

---

*Sears *op. cit.*, p. 453.

**Kerlinger (1964), p. 681.

†Loevinger (1967), p. 93.

The psychometric tradition, which is preserved in many current standardized tests of achievement, emphasizes one correct answer. Kamii (1971) reveals that "these practices reflect an additive view of knowledge and a philosophy of education that values the child's ability to give correct answers."* The "additive view" restricts the probable uses of a response to a single correct response, not taking into consideration that other cultural groups may interpret other uses as being more appropriate for the same response.

An alternative approach which consumes more time in testing is called the exploratory method. This method places emphasis on the process by which the answer is given instead of the end product of the response. The significance of this approach for various cultural groups is that it allows one to have more information with which to evaluate cultural group responses because it answers the question of "why" certain responses appeared. Thus, it becomes easier for the test designer to approach the eventual goal of equivalent statements.

FOURTH: Hypothesize positive and negative relationships between several related behavioral constructs within each culture (variables are operationally defined).

Because of the problems of conceptual equivalence and operational definitions discussed by Sears, there is a definite need for a great deal of testing within cultures before engaging in comparative study between cultures.

Several theoretical variables which have been used to demonstrate performances between groups, as yet, have not been defined sufficiently within groups.

Sears, referring to this difficulty of the interchangeability of behavioral indices, uses this graphic diagram to make his point:

> X → X and Y → Y relationships must be examined carefully before X → Y relationships are sought.**

*Kamii (1971), p. 340.

**Sears (1961), *op. cit.*, p. 447.

At this point, our processes overlap with test theory and test methodological procedures, where the manipulation of test items is based on an investigator's assumptions about reliability and validity.

FIFTH: Identify wide ranges of structural relationships between test items.

Przeworski and Teune demonstrate the mathematical assumptions which can be applied to equivalence.* Their assumptions are applied below to the problem of measuring educational achievement among two cultural groups; in this case, Black and White third grade students.

1. There are a number of items X_1, X_2 .... X_3 that can be used to measure achievement in each cultural group. The assumption that has not been empirically verified is that these achievement items have the same factorial structure between groups and therefore, any subset of these items are also similar. Therefore, the following assumption can be made.
2. A set of item X_k is common to all groups.
3. For each cultural group C_k, there is a set of items X_{N-k} that is specific to the given cultural context of behavioral responses.

Given statements (2) and (3), one may conclude that:

4. For each cultural group C_k, there is a set of test items X that is composed of subsets X_k and X_{N-k}.
5. If items X_1, X_2, X_3 ... X_3 are highly correlated with each other, the set of X_k is considered homogenous. (This is the likely objective of most test items found in traditional objective tests.)
6. However, little empirical data can be found on the homogeneity between subsets of items between groups.

---

*Przeworski and Teune, "Equivalence in Cross-National Research," *The Public Opinion Quarterly*, Vol. XXX, 1966 (p. 551-568).

In other words, for each cultural group, little is known about the intercorrelations of subsets X_k and X_{N-k}, in terms of their structural similarity.

In descriptive comparisons between groups, there has been a greater emphasis on examining the similarity of means and standard deviations. From these scores, inferences have been made about equivalency of measurement statements. As Przeworski and Teune would agree, these judgments have been supported more on the empirical assumptions about the behavior of groups, rather than the actual behavior of the test items within and between these groups. By examining the actual behavior of the test items, one is then free to approach statements about parallel factor distribution patterns.

SIXTH: Establish equivalent measurement statements between two or more cultural groups.

When the following conditions* are met, it is possible to make equivalent measurement statements.

(1) If the analyses show total invariance in the structural patterns of test scores or relationships across cultures, one is able to infer general statement about behavior;

(2) If the analyses reveal partial variance, one is able to infer general statements only from those relationships that are invariant;

(3) If the analyses reveal total variance, one can make general statements only about each culture, but not across cultural groups.

If either conditions (2) or (3), above, are met, it cannot be said that equivalence has been established across groups.

*Refer to the work of: Przeworski and Teune (1970) and Triandis (1972).

As can be seen from the above sections, the paradigm may be applied to the tasks of either making inferential measurement statements from existing tests or serve as a guide in the construction of new tests.

Traditionally, cross-cultural studies have been concerned in part with the degree to which factorial structure is stable among various cultural groups. The question becomes: Why have traditional psychometric practices ignored the lack of empirical psychometric data on the generality of factors between different social groups? The assumption about the similarity of factor structure between and within item clusters is subject to empirical verification; and, until this line of research is exhausted, inferential statements about the performance of various groups may be inaccurate.

Several investigations have documented a list of factors that can produce differences in the structure of abilities:* linguistic systems, genetics, environmental demands and mode of life of subjects. It should be noted that these categories come close to describing the categories found in the literature on culture-fair testing.

It is argued in this paper that the reason for the slow progress in the application of psychometric principles to the structural relationships of group responses, has been almost total reliance of the testing process to a monocultural interpretation. Given a cross-cultural paradigm of testing, the equivalence of cultural groups becomes the fundamental postulate. That postulate does not imply that no differences should exist between groups; but it does imply that, when measurement statements are compared across groups, and when these statements depict patterns of differences between groups, they should have equivalent meaning.

B: POLICY IMPACT OF A PARADIGM SHIFT

Because policymakers in education are under the control of the executive branches of federal and state governments, they are

*Refer to Guthrie (1967), p. 458.

required to view their administrative decisions from a political base. Therefore, the issue of selecting a paradigm for culture-fair testing, as an additional alternative by which to view testing, can be viewed also from a political perspective.

The political problem facing the decision-maker is: How to meet the demands of both minorities and the general public while responding to legislative mandates. Maintenance of this delicate balance can be achieved, in part, through the application of a cross-cultural comparative paradigm for testing.

A cross-cultural paradigm would have the effect of providing some solutions to the dilemmas facing policymakers, as defined in Part I, because the recommended paradigm:

- Suggests a cross-cultural strategy and the establishment of equivalent instruments, permitting the interpretation of valid data within cultures and reliable data across cultural groups.
- Reinforces cultural values among groups by identifying behavior sampling techniques within subject cultural groups as well as among cultural groups.
- Provides a testing alternative in which federal and state decision makers can expand their educational premises on which they formulate policy and legislation.

The evaluation of the effectiveness of federal compensatory educational programs must rely in part, on the results of standardized testing. There is need for some federal policy in the selection and use of tests, especially in programs funded under ESEA Title I, and programs under the EOA Act of 1965.

Currently, many agencies of the Federal government are in the process of developing uniform guidelines for more effective regulation of employee testing.

While the actions of these agencies do not have a direct bearing on the problems of testing in the context of education, they represent the most recent federal initiative in the investigation of

test abuse. The task ahead, therefore, is to explore strategies that could be applied in the evaluation of the selection and use of tests in early childhood and elementary education.

One such effort could involve the cross-cultural alternative presented in this paper. With needed empirical verification, this proposition could be used in the creation of a preliminary process, method or technique that would aid in the effective screening of tests proposed in minority-related research, development and evaluation projects.

SELECTED BIBLIOGRAPHY

Airasian, Peter W. and George F. Madaus, "Criterion-Referenced Testing in the Classroom," in Ralph W. Tyler and Richard M. Wolf (eds.), *Crucial Issues in Testing*, McCutchan Publishing Corp., Berkeley, California, 1974, part 4, Chapt. 8, pp. 73-88.

Anastasi, A., *Differential Psychology*, Macmillan, New York, 1958a.

-------, "Heredity, Environment, and the Question 'How?'," *Psychological Review*, 1958b, 65, pp. 197-208.

Anastasi, Anne, and F. A. Cordova, "Some Effects of Bilingualism upon the the Intelligence Test Performance of Puerto Rican Children in New York City," *Journal of Educational Psychology*, 1953, 44, pp. 1-19.

Anastasi, A. and C. deJesus, "Language Development and Goodenough Draw-a-Man IQ of Puerto Rican Preschool Children in New York City," *Journal of Abnormal Social Psychology*, 1953, 48, pp. 357-366.

Angoff, W. H. and S. F. Ford, "Item-race Interaction on a Test of Scholastic Aptitude," *Journal of Educational Measurement*, 1973, Vol. 10, No. 2, pp. 95-105.

Armstrong, R. A., *Test Bias from the Non-Anglo Viewpoint: A Critical Evaluation of Intelligence Test Items by the Members of Three Cultural Minorities*, Doctoral Dissertation, University of Arizona, 1972.

Berk, Joel S. and Michael W. Kirst, "Federal Aid to Education: Who Benefits? Who Governs?," *Intergovernmental Relations: Conclusions and Recommendations*, Lexington Books, 1972, Chapt. 9, p. 385.

Breland, Hunter M., *An Investigation of Cross-Cultural Stability in Mental Test Items*, Paper from Educational Testing Service, presented at the Annual Meeting of the American Educational Research Association, Chicago, Illinois, April 1974.

Brickman, W. and S. Lehrer (eds.), *Education and the Many Faces of the Disadvantaged: Cultural and Historical Perspectives*, John Wiley and Sons, New York, 1972.

Brigham, C. C., *A Study of Error*, College Entrance Examination Board, New York, 1932.

Burt, C., "The Structure of the Mind: A Review," *British Journal of Psychology*, 1949, 19, pp. 176-199.

Campbell, Donald T. and Albert Erlebacher, "How Regression Artifacts in Quasi-Experimental Evaluations can Mistakenly Make Compensatory Education Look Harmful," in Jerome Hellmuth (ed.), *Disadvantaged Child, Vol. 3*, Brunner/Mazel, New York, 1970, Chapt. 10, pp. 185-210.

Cattell, R. B., "Are IQ Tests Intelligent?," *Psychology Today*, Vol. 1, 1968.

Cicourel, Aaron V. et al., *Language and School Performance*, Ford Foundation, Division of Public Education, Final Research Report, 1970.

Clark, K. B., "The Most Valuable Hidden Resource," *College Board Review*, 29, Spring, 1965, pp. 23-26.

Clasby, Miriam, Maureen Webster and Naomi White, *Laws, Tests and Schooling: Changing Contexts for Educational Decision-making*, Syracuse University Research Corp. Oct. 1973, EPRC Research Report RR-11.

Cleary, T. Anne, *Test Bias: Validity of the Scholastic Aptitude Test for Negro and White Students at Integrated Colleges*, College Entrance Examination Board Research and Development Report No. 18, Princeton, New Jersey: Educational Testing Service, 1966.

Cleary, T. Anne and Thomas L. Hilton, "An Investigation of Item Bias," *Educational and Psychological Measurement*, XXVIII, 1968, pp. 61-75.

Cole, Nancy S., "Bias in Selection," *American College Testing Program*, No. 51, May 1972.

Coleman, James S. et al., *Equality of Educational Opportunity*, Washington, D.C.: U.S. Government Printing Office, 1966.

Daniel, Robert P., "Basic Considerations for Valid Interpretations of Experimental Studies Pertaining to Racial Differences," *The Journal of Educational Psychology*, January 1932, Vol. 23, pp. 15-27.

Darlington, Richard B., "Another Look at 'Cultural Fairness'," *Journal of Educational Measurement*, Vol. 8, 2, 1971 (summer).

Davis, A. and K. Eells, *Davis Eells Games: Davis-Eells Test of General Intelligence or Problem Solving Ability, Manual*, Yonkers-on-Hudson, New York: World Book, 1953.

Dewey, John, "Bureautechnocracy and the Schools," in Charles Tesconi, Jr. and Van C. Morris (eds.), *The Anti-Man Culture*, University of Illinois Press, Urbana, Ill., 1972, p. 36.

Dobzhansky, T., *Genetics and the Origin of Species*, Columbia University Press, New York, 1951.

Dyer H. and E. Rosenthal, "Assessing Education at the State Level: Overview of the Survey Findings," in R. Tyler and R. Wolf *Crucial Issues in Testing*, McCutchan Publishing Corporation, Berkeley, California, 1973, part five, pp. 105-127.

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

Epps, E. G., "Situational Effects in Testing," in LaMar Miller (ed.) *The Testing of Black Students: A Symposium*, Prentiss-Hall, Englewood Cliffs, New Jersey, 1974, Chapt. 5, pp. 41-51.

ETS Developments, "Are Aptitude Tests Unfair to Negroes? ETS Investigates Two Kinds of 'bias'," 1966, 14, 1.

Fein, G. G. and A. Clarke-Stewart, *Day Care in Context*, John Wiley and Sons (Interscience), New York, 1973.

Finley, Carmen J., "Not Just Another Standardized Test," in Ralph W. Tyler and Richard M. Wolf (eds.), *Crucial Issues in Testing*, McCutchan Publishing Corp., Berkeley, California, 1974, Part Five, Chapt. 9, pp. 95-101.

Flaugher, Ronald L., "The New Definitions of Tests Fairness in Selection: Developments and Implications," *Educational Researcher*, A Publication of the American Educational Research Association, Vol. E, No. 9, Oct. 1974, p. 13-16.

Goodlad, John I., "Educational Opportunity: The Context and the Reality," Introduction of *Educational Change: Implications for Measurement*, Proceedings of the 1971 Invitational Conference on Testing Problems, Educational Testing Service, Princeton, New Jersey.

Greene, John, "Why Norm-Referenced," in Lawrence E. Gary (ed.), *Social Research and the Black Community: Selected Issues and Priorities*, Harvard University, Washington, D.C., 1974, Part IV, Sect. 4, pp. 180-186.

Guilford, J. P., "The Structure of Intellect," *Psychological Bulletin*, 1956, 53, pp. 267-293.

Guthrie, G. M., "Structure of Abilities in a Non-Western Culture," in D. Jackson and S. Messick (eds.), *Problems in Human Assessment*, McGraw-Hill, New York, 1967, pp. 458-468.

Havighurst, R., M. Gunther and I. Pratt, "Environment and the Draw-a-Man Test: the Performance of Indian Children," *Journal of Abnormal Social Psychology*, 1946, 41, pp. 50-63.

Hieronymus, A. N., "Today's Testing: What Do We Know How to Do?," *Educational Change: Implications for Measurement*, Proceedings of the 1971 Invitational Conference on Testing Problems, Educational Testing Service, Princeton, New Jersey, pp. 57-68.

Hirsch, J., "Behavior-Genetic Analysis and Its Biosocial Consequences," in Miller and Dreger (eds.), *Comparative Studies of Blacks and Whites in the United States*, Seminar Press, New York, 1973, pp. 34-51.

Hoffmann, B., *The Tyranny of Testing*, Crowell-Collier Press, New York, 1962.

Huff, Sheila, "Credentialing by Tests or by Degrees: Title VII of the Civil Rights Act and Griggs V. Duke Power Company," *Harvard Educational Review*, Vol. 44, No. 2, May 1974, pp. 246-269.

Humphreys, L. G., "The Organization of Human Abilities," *American Psychologist*, 1962, 17, pp. 475-483.

Inhelder, B. and J. Piaget, *The Early Growth of Logic in the Child*, 1974, translated by E. Lunzer and D. Papert, Norton Inc., New York, 1969.

Jencks, Christopher et al., *Inequality, A Reassessment of the Effect of Family and Schooling in America*, Basic Books, Inc., Publishers, New York and London, 1972, Chapt. Three, Inequality in Cognitive Skills, pp. 52-109; Chapt. Five, Inequality in Educational Attainment, pp. 135-158.

Jensen, A. B., "Another Look at Culture-Fair Testing," in *Western Regional Conference on Testing Problems*, 1968 ETS proceedings, ETS, Princeton, New Jersey, May 3, 1968.

Kamii, C., "Evaluation of Learning in Preschool Education," in B. Bloom, J. T. Hastings and G. Madaus (eds.), *Handbook on Formative and Summative Evaluation of Student Learning*, McGraw-Hill, New York, 1971, Chapt. 13, pp. 281-344.

Kerlinger, Fred N., *Foundations of Behavioral Research*, Holt, Rinehart and Winston, Inc., 1964, Chapt. 9: Purpose, Approach and Method, p. 147; Chapt. 23: Foundations of Measurement, p. 411; Chapt. 32: The Semantic Differential, p. 564.

Kimura, D., "The Asymmetry of the Human Brain," *Scientific American*, 1973, 228(3), March, pp. 70-78.

Kirp, D., "Student Classification, Public Policy, and the Courts," *Harvard Educational Review*, Vol. 44, No. 1, February 1974, pp. 7-52.

Klineberg, O., "An Experimental Study of Speed and Other Factors in Racial Differences," *Archives of Psychology*, No. 93, 1928.

-------, *Race Differences*, Harper, New York, 1935.

Klitgaard, Robert E., *Achievement Scores and Educational Objectives*, The Rand Corp., R-1217-NIE, January 1974.

Kuhn, Thomas S., *The Structure of Scientific Revolutions*, The University of Chicago Press, Vol. II, No. 2, 1962.

L'Abate, Luciano, Yvonne Oslin, and Vernon W. Stone, "Educational Achievement," in Kent S. Miller and Ralph Mason Dreger (eds.), *Comparative Studies of Blacks and Whites in the United States*, Seminar Press, New York and London, 1973, Chapt. 11, pp. 325-354.

Lawrence, P. J., "Symposium: A Study of Cognitive Error Through an Analysis of Intelligence Test Errors," *British Journal of Educational Psychology*, 1957, Vol. 27, pp. 176-181.

Lazersfeld, P. F. and Morris Rosenberg, *The Language of Social Research*, Free Press, Glencoe, Ill., 1955, (sect. IV).

Lesser, G. G. Fifer, and D. Clark, *Mental Abilities of Children in Different Social and Cultural Groups*, Cooperative Research Project No. 1635, 1964.

Lester, David, "The Subject as a Source of Bias in Psychological Research," *Journal of General Psychology*, Vol. 81, 1969, pp. 237-248.

Levine, D., "Educational Alternatives for the Disadvantaged Child," in Brickman and S. Lehrer (eds.), *Education and the Many Faces of the Disadvantaged: Cultural and Historical Perspectives*, John Wiley and Sons, New York, 1972, p. 42.

Linn, Robert and Charles E. Werts, "Considerations for Studies of Test Bias," *Journal of Educational Measurement*, Vol. 8, No. 1, Spring 1971, pp. 1-4.

Loevinger, Jane, "Objective Tests as Instruments of Psychological Theory," in Douglas N. Jackson and Samuel Messick (eds.), *Problems in Human Assessment*, McGraw-Hill, New York, 1967, Chapt. 5, pp. 78-123.

Maccoby, E. E., "Sex Differences in Intellectual Functioning," in E. Maccoby (ed.), *The Development of Sex Differences*, Stanford University Press, Stanford, California, 1966.

MacKay, Robert, "Standardized Tests: Objective/Objectified Measures of 'Competence'," in Cicourel, *Language and School Performance*, Ford Foundation, Division of Public Education, Final Research Report, 1970.

Manning, W. H., "The Measurement of Intellectual Capacity and Performance," *Journal of Negro Education*, 1968, Vol. 37, pp. 258-267.

Mayeske, George, *Technical Paper No. 1*, Office of Program Planning and Evaluation, 1969.

Meier, Deborah, *Reading Failure and the Tests*, New York: Workshop Center for Open Education, 1973.

Mercer, Jane, "A Policy Statement on Assessment Procedures and the Rights of Children," *Harvard Educational Review*, Vol. 44, No. 1, 1974a, pp. 125-141.

-------, *The Who, Why, and How of Mainstreaming*, paper presented for Joint Education-Psychology Division Luncheon, American Association on Mental Deficiency (National Convention), June 3-7, 1974b.

Miller, K. and R. Dreger, *Comparative Studies of Blacks and Whites in the United States*, Seminar Press, New York, 1973.

Mosteller, Frederick and Daniel P. Moynihan, *On Equality of Educational Opportunity*, Vintage Books, New York, 1972.

National Education Association, *Conference Report on Testing, 1972*, adopted by the NEA Task Force in May 1973; Adopted as a resolution by the NEA in July, 1974.

Porteus, S. D., "Racial Group Differences in Mentality," *Tabul. Biol., Haag*, 1939, 18, pp. 66-75.

-------, *The Porteus Maze Test and Intelligence*, Pacific Books, Palo Alto, California, 1950.

Porteus, S. D., and M. E. Babcock, *Temperament and Race*, Gorham, Boston, 1926.

Przeworski, Adam and Henry Teune, "Equivalence in Cross-National Research," *The Public Opinion Quarterly*, Vol. XXX, 1966, pp. 551-568.

-------, *The Logic of Comparative Social Inquiry*, John Wiley and Sons, Inc., New York, 1970.

Raven, J. C., *Guide to Using the Coloured Progressive Matrices*, Lewis, London, 1956.

Roberts, Elsa, *An Evaluation of Standardized Tests as Tools for the Measurement of Language Development*, U.S. Department of Health, Education and Welfare, Office of Education, May 1970, 17p.

Roman, F. W., *The New Education in Europe*, E. P. Dutton and Co., New York, 1930.

Rosenthal, R., *Experimenter Effects in Behavioral Research*, Appleton-Century-Crofts, New York, 1966.

Rychlak, Joseph F., *Foundations For a Logical Learning Theory*, An unpublished paper from Purdue University, 1973.

Sattler, J. M., "Racial 'Experimenter Effects' in Experimentation, Testing, Interviewing, and Psychotherapy," *Psychological Bulletin*, 1970, 73, pp. 137-160.

Sears, Robert, "Transcultural Variables and Conceptual Equivalence," in Bert Kaplan (ed.), *Studying Personality Cross-Culturally*, Row Peterson, Evanston, Ill., 1961, p. 453.

Shuey, A., *The Testing of Negro Intelligence*, J. P. Bell, Lynchburg, Va., 1958.

Spearman, C., *The Abilities of Man*, Macmillan, New York, 1927.

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

Terman, L. M., *The Measurement of Intelligence*, Houghton Mifflin, Boston, 1916.

Terman, L. M. and L. E. Tyler, "Psychological Sex Differences," in L. Carmichael (ed.), *Manual of Child Psychology*, Wiley, New York, 1954, pp. 1064-1114.

Thorndike, Robert L., "Concepts of Culture-Fairness," *Journal of Educational Measurement*, Vol. 8, 2, 1971 (summer).

Thurstone, L. L., *Primary Mental Abilities*, University of Chicago Press, Chicago, 1938.

Timpane, Michael, "Educational Experimentation in National Social Policy," *Harvard Educational Review*, Vol. 40, No. 4, November 1970.

Triandis, H., *The Analysis of Subjective Culture*, Wiley-Interscience, New York, 1972.

Vernon, P. E., "Environmental Handicaps and Intellectual Development: Part I and Part II," *British Journal of Educational Psychology*, 1965, 35, pp. 1-22.

-------, *The Structure of Human Abilities*, Methuen, London, 1960.

Whittke, Carl, "Historical Background: Immigration Policy Prior to World War I," in Benjamin M. Ziegler (ed.), *Immigration: An American Dilemma*, D. C. Heath and Co., Boston, 1953, pp. 1-10.

Williams, Robert L., "Abuses and Misuses in Testing Black Children," reprinted from *The Counseling Psychologist*, Vol. 2, No. 3, 1971.

-------, "Black Pride, Academic Relevance, and Individual Achievement," *The Counseling Psychologist*, Vol. 2, No. I, 1970, pp. 18-22.

Wolfram, Walt, "Levels of Sociolinguistic Bias in Testing," *Seminar in Black English*, Erbaum Publishers, April, 1974.

Ziegler, Benjamin M., *Immigration: An American Dilemma*, D. C. Heath and Co., Boston, 1953.

Zirkel, Perry Alan, "Spanish-Speaking Students and Standardized Tests," *Urban Review*, June 1972, pp. 32-40.