

To the Graduate Council:

I am submitting herewith a dissertation written by Stephen Leigh Tench entitled "A Comparison of Three Intelligence Tests as Non-Biased Measures of Academic Potential." I have examined the final copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Psychology.

Richard A. Saudargas
Richard A. Saudargas, Major Professor

We have read this dissertation
and recommend its acceptance:

J. Albert Tuberville
Michael G. Johnson
Charles L. Thoyson

Accepted for the Council:

Cowminkel
The Graduate School

A COMPARISON OF THREE INTELLIGENCE TESTS AS NON-BIASED
MEASURES OF ACADEMIC POTENTIAL

A Dissertation

Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Stephen Leigh Tench

August 1984

ACKNOWLEDGEMENTS

Having reached this stage of my professional career, I must stop and express my appreciation to several persons who have been outstanding in their support and by their example.

I wish to thank Dr. Richard Saudargas for doing so well what he modestly claims is "what I am paid to do"--even on weekends and holidays.

I was also helped by Mr. Vernon Coffey, Superintendent of Grainger County Schools, who allowed me extraordinary flexibility in my work schedule to pursue my doctoral studies, and Dr. Lorraine A. Hart and Ms. Kathy Self, who kindly helped me in establishing reliability of scoring.

I must also mention the names of three people who have been my models in the ongoing task of integrating professionalism and personal integrity, and who, for some reason, extended to me their friendship: Dr. Jim Posner, Dr. J. Albert Wiberley and Mr. Robert Tipps.

Finally, I come to my my wife, Jean, and my daughter, Sara. I hope they know how much I appreciate them. If not, I am a failure regardless of how many academic degrees or honors I might collect during my lifetime.

ABSTRACT

This study investigated the relationship among three intelligence tests—the Wechsler Intelligence Scale for Children--Revised (WISC-R), the System of Multicultural Pluralistic Assessment (SOMPA) and the Kaufman Assessment Battery for Children (K-ABC)—and their relationship with concurrent and future reading and arithmetic achievement as measured by the Peabody Individual Achievement Test (PIAT) and the Wide Range Achievement Test (WRAT). This relationship is significant because the theoretical bases of the SOMPA and K-ABC dictate that as more pure measures of academic potential, they should not reflect students' culture, socioeconomic status or urban/rural setting to the degree that the WISC-R does.

Subjects included forty rural, low-income white students enrolled in a small school system in eastern Tennessee. Subjects were all recipients of free school lunches under income requirements established by the United States Department of Agriculture. Numbers of subjects were split evenly according to sex and age (6- or 11-years-old). The latter was the maximum age span which would allow overlap of the three intelligence tests.

The full scale scores obtained from the three intelligence tests were significantly correlated ($p < .01$, adjusted for repeated tests of significance). This suggested that subjects received similar patterns of scores from the three tests.

The remaining analyses of data were conducted by ANOVA procedures which accommodated the use of repeated measures from the subjects.

Critical ratios for F were also adjusted for repeated testing for significance.

There were no significant differences with $p < .05$ found in the sample's mean scores obtained from the three intelligence tests, suggesting that the SOMPA and K-ABC were not discernably less influenced by the subjects' families' income and rural setting than was the WISC-R. Also no effects due to the factors of sex, age, or to interaction of factors was found.

Many state departments of education require that determinations of the existence of learning disabilities be made using an achievement-intelligence discrepancy model which assumes that "normal" achievement standard scores should equal the intelligence standard score. Using data values of $[\text{Obtained Achievement}]/[\text{Expected Achievement}]$ where Expected Achievement was replaced by each intelligence test's full scale standard score, analyses were made to determine whether one of the three tests was a better predictor of either current or future achievement. No differences in accuracy were found with respect to concurrent or predictive (7 months later) achievement measures; PIAT or WRAT scores; and reading decoding, comprehension or arithmetic computation subtest scores.

The general conclusion was that the three intelligence tests should lead to similar global findings when used with students from a population similar to that of the subjects. The tests did not yield results appreciably different from each other with respect to accuracy in predicting achievement.

The finding of equal means for the three sets of intelligence test scores was inconsistent with previous findings which reported a 5 point SOMPA--WISC-R discrepancy and, with one exception, a 3 to 4 point K-ABC--WISC-R discrepancy for similar subjects. It was suggested that a sample from even lower income or comprised of nonwhite subjects would be a better test of the SOMPA's performance. The present sample was possibly too close to average with respect to culture and living conditions to represent the groups for which the SOMPA was originally intended. There did not seem to be common characteristics among the K-ABC studies to allow suggestions to be made regarding why the present findings were discrepant.

TABLE OF CONTENTS

SECTION	PAGE
I. INTRODUCTION	1
Statement of Purpose	1
Intelligence--Implications of the Term	2
What is Intelligence?	3
Cultural Retardation	5
Social and Familial Correlates to Measured Intelligence	13
Bias and Its Meaning(s)	14
How the Various Definitions Interrelate	21
Evaluation of Current Tests According to Definitions of Bias	23
Cultural Pluralism and Mental Testing	29
Important Assumptions Underlying Intelligence Testing	31
Standardization	31
Consequences of Inappropriate Testing	41
Labeling	44
Recommendations for Changes in Evaluation Process	48
Calls for Moratorium of Testing	52
Culture-Fair Testing Movement	54
Culture-Specific (Pluralistic) Tests and Adjustment of Norms	57
Piagetian Assessment Instruments	62
Changes in the Clinical Testing Interview: The Examiner and the Groundrules	63
Criterion Referenced Tests	66
Development of New Types of Tests	67
Use of Learning Instruments	68
Narrowed Criteria for Mental Retardation Label	71

SECTION	PAGE
II. METHOD	74
Subjects and Setting	74
Instruments	76
Procedure	81
Data Analysis	82
Hypotheses	83
III. RESULTS	86
Reliability	86
Comparison of Mean Ability Scores	86
Comparison of Distribution of Scores	88
Accuracy in Prediction	88
IV. DISCUSSION	105
LIST OF REFERENCES	111
VITA	121

LIST OF TABLES

TABLE	PAGE
1. Prevalence of IQs below 75 as a function of social class . .	10
2. Grade placement for subjects in sample	75
3. Summary of age data for subjects	75
4. Degree to which sample saturated population of interest at each school	77
5. Achievement data collected according to sample	84
6. Means, standard deviations and ranges of WISC-R, SOMPA, K-ABC and PIAT scores	87
7. Summary of analysis of variance for sex x age x test WISC-R Full Scale IQ, SOMPA Full Scale ELP and K-ABC MPC) . .	89
8. Pearson product-moment correlations among the WISC-R, SOMPA and K-ABC global scores	90
9. Correlations among WISC-R, SOMPA, K-ABC and PIAT	91
10. Correlations among WISC-R, SOMPA, and K-ABC full scale scores vs. achievement test scores for 6-year-old sample . . .	92
11. Correlations among WISC-R, SOMPA and K-ABC full scale scores vs. achievement test scores for 11-year-old sample . .	93
12. Summary of analysis of variance for sex x age x ability tests' concurrent prediction of K-ABC Reading Decoding subtest scores	95
13. Summary of analysis of variance for sex x age x ability tests' concurrent prediction of PIAT Reading Recognition subtest scores	95
14. Summary of analysis of variance for sex x ability tests' concurrent prediction of WRAT Reading subtest scores (6-year-old data only)	96
15. Summary of analysis of variance for sex x ability tests' concurrent prediction of WRAT Reading subtest scores (GE scores; 11-year-old data only)	96
16. Summary of analysis of variance for sex x ability tests' concurrent prediction of K-ABC Reading Comprehension subtest score (11-year-old data only)	97

TABLE

PAGE

17.	Summary of analysis of variance for sex x age x ability tests' concurrent prediction of K-ABC Arithmetic subtest scores	97
18.	Summary of analysis of variance for sex x age x ability tests' concurrent prediction of PIAT Arithmetic subtest scores	98
19.	Summary of analysis of variance for sex x ability tests' concurrent prediction of WRAT Arithmetic subtest scores (6-year-old data only)	98
20.	Summary of analysis of variance for sex x ability tests' concurrent prediction of WRAT Arithmetic (GE) subtest scores (11-year-old data only)	99
21.	Summary of analysis of variance for sex x age x ability tests' future prediction of PIAT Reading Recognition subtest scores	102
22.	Summary of analysis of variance for sex x age x ability tests' future prediction of PIAT Reading Comprehension subtest scores	102
23.	Summary of analysis of variance for sex x ability tests' future prediction of WRAT Reading subtest scores (6-year-old data only)	103
24.	Summary of analysis of variance for sex x age x ability tests' future prediction of PIAT Arithmetic subtest scores . .	103
25.	Summary of analysis of variance for sex x ability tests' future prediction of WRAT Arithmetic subtest scores (6-year-old data only)	104
26.	Summary of analysis of variance for sex x ability tests' future prediction of WRAT Arithmetic (GE) subtest scores (11-year-old data only)	104

I. INTRODUCTION

Encouraged by federal legislation and ensuing regulations plus emerging case law, considerable attention has been focused on the need for educational assessment instruments that do not discriminate against minorities. The System of Multicultural Pluralistic Assessment or SOMPA (Mercer, 1978) has been proposed as a nondiscriminatory assessment battery that yields a statistic of "estimated learning potential." The ELP has been advanced as offering a more nearly "true" measure of learning potential than does the traditional deviation IQ from the Wechsler Intelligence Scale for Children--Revised (Wechsler, 1974). The ELP is derived from the Wechsler IQ by statistical procedures that "partial out" such "contaminating" factors as socioeconomic status, acculturation, minority status and family size.

A second evaluation instrument, the Kaufman Assessment Battery for Children (Kaufman & Kaufman, 1983a) has recently been published and promoted as a superior measure of learning potential for disadvantaged and minority youth because the scale more carefully separates achievement variables from what are considered to be more purely learning skill variables. The statistic of general learning potential derived from this test is called the mental processing composite or MPC.

Statement of Purpose

The purpose of the present study is to examine the above three assessment instruments--the WISC-R, SOMPA and K-ABC--with respect to their relative efficacy in predicting concurrent and future academic

attainment of an elementary school population that is poor, white and rural.

In evaluating a new assessment device, psychologists are frequently interested in how the test performs with respect to the variables of sex, ethnic group, socioeconomic status, age at assessment and IQ range. The present experiment's two-by-two within-subjects design controlled for the variables of sex and age. (The age range considered in this study is that age span for which the norms of the three instruments overlap.) The use of a free-lunch eligibility criterion assured that the lower socioeconomic class was addressed. This is appropriate since the K-ABC and SOMPA are designed to minimize the effects of cultural and economic disadvantage. However, it should be noted that no racial or ethnic minority groups were represented in the subject sample.

Intelligence--Implications of the Term

There are measured differences between the middle-class white and minority groups on measures of intellectual ability. Why should these well-documented differences inspire the vehemence that characterize their discussion? The answer is both obvious and pervasive. The lay public generally perceives "intelligence" to be an all-embracing capacity to achieve at any task, to be finite and predetermined at birth, and to be only marginally susceptible to environmental manipulation. "Intelligence tests" were born from these assumptions and are still widely understood to measure the concept as perceived.

Turnbull in Samuda (1975) pointed out that as long as the definition of "intelligence" is not standardized, the logical result

is controversy. The significance of reported differences in test scores has been compounded by several developments: the competition for choice occupations increasingly has depended upon intelligence and related measures, minorities increasing sense of pride and self-awareness has made these groups much less willing to be evaluated by middle-class Anglo standards, and the recurrence of the heritability issue.

What is Intelligence?

"Intelligence" in the currently understood sense is probably a relatively new concept in human language. Prior to the 20th century, determination of an individual's mental ability would likely have involved assessing mastery of the bulk of information and skills helpful to survival in his or her culture. A thirty-year-old who knew twice as much as another person would be considered twice as able as the other person, regardless of whether the other person were thirty, twenty, or ten. After the development of the IQ concept, however, no longer did society evaluate the individual's mental ability against a scale which was cumulative and comprised of things which society had determined were important to master. With the transformation of the ability concept from that of accumulated skills and knowledge to rate of acquisition of skills and knowledge, a ten-year-old who had accumulated a great deal less of the society's valued knowledge than a thirty-year-old could be considered "more intelligent" than the older person because the youngster knew more than was expected for someone his or her age. Thus mental ability changed from an achievement-based concept to a psychometric concept.

What is "intelligence"? The answer is of extreme importance as the definition greatly influences how it is measured, what may be said about it, and what conclusions may be drawn about the person whose "intelligence" is being measured. The offerings are many, and each stresses certain operations and skills (Pyle, 1979). Binet viewed intelligence as the ability to judge well, to comprehend well and to reason well. Spearman looked for a general intelligence which involved mainly the "education of relations and correlates." Terman saw intelligence as the capacity to form concepts and to grasp their significance. Burt wrote about innate, general, cognitive ability. For Heim, intelligence was grasping the essentials of a situation and responding appropriately. Wechsler sought for the "aggregate or global capacity of the individual to act purposefully, to think rationally and to deal effectively with the environment." For Piaget, the growth of intelligence involved increased adaptability to the physical and social demands of the environment. At the operationalist end of the spectrum, Boring defined intelligence as "what the test measures."

The study of intelligence also differs according to the basis of the model (Pyle, 1979). Psychometric or test-based models of intelligence lend themselves to manipulation and correlational studies. The two factor models of Spearman (g and s), Cattell ("fluid" and "crystallized" intelligence), and Jensen (Level I associative and Level II cognitive intelligence), the hierarchical models of Burt and Vernon, and the multivariate models of Thurstone and Guilford are examples. The developmental models of Piaget and Hebb lend themselves more to

long term study of psychological processes that lead to intelligent behavior.

Cultural Retardation

Subnormal intelligence is a concept that is anchored in compiled data collected about the developmental rates of a selected sample of culturally-valued skills. While there is little consensus in terms (Sarason and Doris, 1969), subnormal intellectual functioning is frequently divided into two conceptual subgroups by psychologists: mental retardation and mental deficiency. Mental retardation refers to intellectual performance less than that normally expected due to social and environmental factors. This concept may be considered to reject the medical model assumption implicit in the term mental deficiency, which includes the observed deficits of intellectual performance but adds observed deficits of environmental adaptation and assumes an organic base to the observed deficits.

The lack of consensus has several contributing factors including differing terminology employed by psychologists and educators and the outside imposition of legal terminology from jurists, legislators and writers of regulations who frequently fail to recognize, gloss over, or choose to ignore the semantic differences of terms used by their expert witnesses and consultants. This state of affairs obviously leads to problems in communication. In this study, when reference is made to mental retardation or mental deficiency, the terms will be used in the different modes defined in the preceeding paragraph.

There is an additional semantic problem when using the term "intelligence." Symptomatic of the semantic problem is Wolman's

(1973) Dictionary of Behavioral Science listing of six theoretical definitions of intelligence corresponding to six prominent theorists' conceptualizations of measurement schemata. Another popular notion of "intelligence" is to endorse Boring's (1923) dictum that intelligence is, by definition, that which intelligence tests measure. To clarify the language in this study, the term "intelligence" is used to refer to Spearman's g , that is, that hypothetical source of variance which is used to explain the covariance of scores across "intelligence" tests measuring intelligence from various conceptual orientations (Spearman, 1904). Sidestepping the argument regarding whether g is a unitary ability or a conglomerate of many subordinate abilities, intelligence (" g ") is referred to in a statistical sense as the source of individual differences in scores which is common to a number of different tests (Jensen, 1969a, p. 9).

Intelligence, as used here, is measured imperfectly by intelligence tests in common use. It is pointless to attempt to state what the loading of g is, for example, on the WISC-R or the Binet. Nevertheless, when the intercorrelations among a dozen or more Wechsler- and Binet-type tests are factor analyzed, the results typically indicate that g accounts for 50 percent or more of the total individual differences in variance in all the tests (Jensen, 1969a, p. 10). Thus when the term "measured intelligence" is used, it refers to scores on "intelligence tests."

As is well known, the first modern test of intelligence came into existence to help select French school children who would likely have to be placed eventually into schools for the mentally deficient. To

this date, intelligence tests are primarily used for purposes of prediction of success in academic settings. That predictability does not in itself account for the popularity of intelligence testing, because if one were only interested in the next year's school grades, they could be predicted much better from this year's grades than from an intelligence test score (Cattell, 1968). In the present educational setting, intelligence testing is performed because the process does predict academic success, purports to provide information which helps to understand the reasons behind an individual's success or failure in school, and because it is generally believed that intelligence test scores can predict performance in important out-of-school situations (for example, productivity in certain job situations).

Some critics feel that the concept of intelligence itself is theoretically misleading. For example, Kagan (1973) calls for separation of the measurement of basic cognitive processes (memory, inference and so on) from measurement of acquired knowledge, and a recognition that competence attainment and knowledge acquisition are "totally dependent" upon the demands placed by the testee's culture and opportunities which it offers. Kagan insists that intelligence must be conceptualized relative to the testee's culture. From the opposing viewpoint, Carroll and Horn (1981) criticize intelligence tests for being too culturally specific. This lack of a consensus about theoretical underpinnings of intellectual assessment leads to the situation where the same score may be gotten by use of many different types of skills.

Most tests of intelligence are multiple-choice group tests, which are given to large numbers of school children every year in educational settings across the Western world. Individually administered tests of intelligence are generally considered to be more accurate estimators of intelligence because the combination of undivided attention of the examiner and greater versatility of the instrument lends itself to avoidance of many potentially invalidating influences such as poor motivation, lack of attention and misunderstanding of directions. However, when the topic of intelligence testing is discussed, even in the professional literature, only infrequent discrimination is made between the group and individual testing paradigms.

The IQ score which is derived from most intelligence tests is an index of the rate of cognitive development and rate of learning. Psychometrically the IQ score is a robust measure of academic potential. IQ scores and school achievement measures consistently correlate with r 's in the .50's, higher at the elementary level and falling somewhat into the college years (Brody & Brody, 1976). Critics point out that this predictive ability is of little practical or theoretical interest since schools already have the achievement test scores, and the correlations may be spuriously raised because they are measured by such similar methods as to be practically identical (Garcia, 1981; Findley, 1975). IQ tests also correlate well with number of years of education eventually completed and success in completing college, yet a large percentage of college students with IQs below 110 complete college (Brody & Brody, 1976; Strauss, 1981).

Correlations between summed scores of Piagetian tests and Binet IQs are in the .70's for children of a given age. Piaget felt that IQ tests measured the same concept as Piagetian tests, albeit in less pure and conceptually understandable form (Kohlberg, 1968).

Intelligence testing becomes quite controversial when its practice becomes equated with the perpetuation of social inequalities which have been attacked by much effort and social legislation. The correlations between social status and IQ ranges between .50 and .70, and between parents' occupation and children's IQ about half this. That IQ scores discriminate so well between the "haves" and "have nots" in our society could scarcely have been overlooked by advocates of equal educational opportunity. A telling indicator is the estimated prevalence of children with IQs below 75 as a function of social class and racial group, reported as percentages in Table 1.

Jensen (1970) remarks that it is extremely difficult to find subjects from the IQ range from about 50 to 85 in the middle or upper-middle socioeconomic classes, and that it is much easier to locate gifted children (i.e., $IQ > 130$) in lower socioeconomic classes than to find the mildly retarded in the upper class. He cites the Scottish National Survey (Maxwell, 1953) which established that high intellectual ability is more evenly dispersed across different social classes than is low intellectual ability. There is no ceiling for intellectual ability for the various occupational groups, while there is a steadily rising lower bound as one ascends the occupational status ladder.

Table 1. Prevalence of IQs below 75 as a function of social class^a

Socioeconomic Status	White	Black
High 1	0.5	3.1
2	0.8	14.5
3	2.1	22.8
4	3.1	37.8
Low 5	7.8	42.9

^aReported as percentages. This data is taken from Jensen (1970).

Actual IQ distributions arising from intelligence tests tend to be flatter than the normal curve with fewer extreme scores. The IQ distribution curve is actually two superimposed curves, one a "normal" distribution and one with far fewer cases, comprised of severe mental deficiency accompanied by physical abnormalities and/or clear signs of neurological damage. The second distribution has for the most part IQs below 50, and accounts for an incidence of IQs less than 45 that is 18 times the expected incidence (Jensen, 1970).

The very low IQ retardates (below about 40) are generally reproductive casualties and are fairly well distributed across socio-economic groups (Jensen, 1970; Isaacson & von Hartesveldt, 1978). Siblings of severely subnormal retardates have average IQs whereas siblings of familial-cultural retardates show the expected regression toward the mean. The overrepresentation of retardates in the lower classes is found in the familial-cultural group which corresponds generally to the school group diagnosed as "educable mentally retarded," and is actuarially confounded with ethnic/cultural factors. Such a classification is normally made when IQ ranges between 50 and 70, there are only minor adaptive behavior deficits, and there is no known organic basis for the retardation (Jensen, 1970).

This increased incidence of mental retardation in the lower classes is frequently assigned to the culturally and intellectually impoverished environment which characterizes many lower SES homes. This does not satisfactorily explain, however, how fewer than one-half of the children of retarded parents are ever diagnosed as themselves retarded. Of children diagnosed as retarded, 48.3 percent have at

least one parent with $IQ < 70$. Among families with both parents with $IQ > 70$, the incidence of mentally retarded offspring is 0.5 percent. Among families with one or more parent with $IQ < 70$, or with one or more aunt or uncle with $IQ < 70$, the incidence of mentally retarded offspring rises to 2.5 percent (Reed & Reed, 1965, cited in Jensen, 1970).

Biological factors affecting the IQ of children include maternal age, maternal diet particularly during the last few months of pregnancy, the child's diet especially during the first two years, and maternal smoking, drinking, drug use, disease, complications of delivery and term of pregnancy. Sociocultural factors which account for substantial amounts of variance in IQ scores (Pyle, 1979) include low maternal IQ, SES, birth order and spacing, years of schooling of the child, and living in isolated mountain communities (Hirsch, 1928; Sherman & Key, 1932; Wheeler, 1932; Chapanis & Williams, 1945). The environment has the most damaging effect on a developing organism during periods of most rapid growth. For the human brain, this is from the second trimester of pregnancy to the end of the second year. Malnutrition and dietary inadequacies are generally accompanied by poverty, lack of education and medical problems. Lack of stimulation which characterizes many lower SES homes is also frequently related to lower scores on intellectual measures. Animal studies show that early stimulation-impooverished environments lead to lower brain weight, decreased thickness of cortex, decreased complexity of cortical cells and behavioral changes. Yet other animal studies show that the effects of early malnutrition, thyroid deficiency, brain lesions and

poor stimulation can be ameliorated by "superenriched" environments (Isaacson & von Hartesveldt, 1978).

Thorndike and Hagen (1977) identified six elements common to environments which provide poor background for performing well on intelligence tests: (1) language used at home is not standard English, (2) academic achievement is not allocated emphasis, (3) home environment is unstable and unsupportive because of social and economic pressures, (4) parents do not function as "teachers" to their young children, (5) toys, books and games are in short supply, and (6) the family feels alienated from the dominant culture.

Social and Familial Correlates to Measured Intelligence

Perhaps the best known correlate to intelligence is social class or socioeconomic status (SES). In a review of the literature, Jensen (1969a, p. 75) pointed out that by school age, children's IQs are substantially correlated with the SES of their parents. This finding holds true across the world's cultures and the research extends to the beginning of the century. Jensen reported that half of all the reported correlations between SES and children's IQs fell between .25 and .50, with most falling between .35 and .40. Furthermore, the mean IQs of social classes vary over a range of 15 to 30 points depending upon the method of status classification.

A hot debate has raged in the literature (and behind university lecterns and in the popular media) over the source of the measured IQ differences between the various ethnic groups in our society. While both sides of this issue tend to agree that the necessary studies have not taken place to properly test the nature of the score differences

and their meaningfulness, they vehemently disagree over more fundamental issues. These issues include whether intelligence should be conceptualized as including a component which is correlated with traditionally accepted ideas of successful adaptation to the larger culture (e.g., income, scholastic attainments), to the even more basic issue of whether it is ethically responsible to even consider such inflammatory ideas as hypothesized ethnic differences in intellectual potential.

Both those critics opposed to and those favoring research into the origins of ethnic test-score differences exemplify the dictum that questions asked of data are dictated by the ontological world view of the observer. Opponents such as The Observer of Psychological Record (1970) point out that the vernacular use of the concept "intelligence" confuses actions (i.e., the behavioral reactions to specific situations reflecting the inseparable interaction of environment, nature, history and so on) with things which are ultimately genetically based and therefore transmittable. Thus from this point of view the question of ethnic differences in intelligence continue to arise even though the question is ultimately meaningless as it is unanswerable within the confines of psychological linguistic and philosophical postulates. Similar themes have been expressed by others (e.g., The Observer, 1973; Schoenfeld, 1974).

Bias and Its Meaning(s)

In coming to an understanding of this issue, one should begin by defining the terms for discussion and then distinguishing between bias as a psychometric statistic and bias as a negative behavioral

phenomenon which has deleterious effects upon its object. That is the purpose of the following paragraphs.

Psychological tests do discriminate among people--that is their function. Psychological tests are designed to magnify individual differences, and it would be surprising if these tests failed also to display differences between groups with fairly similar genotypes and environmental backgrounds. However, as already stated in the introduction, the presence of these known group differences leads to a furor when an interpretation is made as to their meaning.

There currently seems to be a fairly broad consensus within the testing literature that a meaningful distinction drawn between aptitude tests and achievement tests, per se, is outmoded, since the primary difference is that an "aptitude" test is constructed to contain more items measuring general information than the more specific content of "achievement" tests. In reality, all aptitude tests are achievement tests, simply because one cannot measure what has not been acquired and thus produced in a measurable format. Nevertheless, a key concept in interpretation of test scores is the conceptual distinction between aptitude and achievement testing.

For example, if a class of students obtains high marks on an achievement test, then society congratulates its teachers as having taught a course of study well. If the class does poorly on the achievement test, then pressure is brought to bear on the teachers and the students to apply themselves again to the material to be mastered. On the other hand, if the test is perceived to be a measure of aptitude, the interpretation is quite different. The class which does

well on the test is widely perceived as capable on all academic tasks or even all major life tasks. The students who do poorly on the "aptitude" test are seen as comprehensively inadequate. The societal implications are vast:

If test performance is low and it is seen as an index of achievement, then there is pressure to increase the application of society's educational resources to improve that achievement. If the test result is seen as an index of aptitude, however, then the same low test score may be seen as a justification of the withdrawal of educational resources. In the one case, society assumes greater responsibility; in the other, it escapes or denies responsibility. (Flaughner, 1978)

Critics have pointed to test results as harmful evidence used by components of society not wanting to improve opportunities for disadvantaged classes. Low test scores may be used both as justification of current status and rationale for restriction of future opportunity (Pyle, 1979; Gordon & Terrell, 1981).

Failure to clearly recognize the conceptual distinction between achievement testing and aptitude testing when interpreting test results has led to great emotionality and what may be characterized as avoidance behavior on the part of some political leaders. For example, instead of demands to "fix the system" to raise minority student achievement, there have been many calls for the abandonment of "lying" tests, which are seen as indicating that minority students are intellectually incapable. An important distinction also needs to be drawn between what is termed by Clarizio (1978, 1979) as cultural loading and cultural bias. Cultural loading is determined by the specificity of the test content as selection is made from the domain of possible test items, to yield those items which eventually are found in the test. As an example, the item "Name the last two

governors of Michigan" is more culturally loaded than "Name the last two presidents of the United States." Cultural bias is an empirical finding of differential predictive validity and differential internal test characteristics. Clarizio posits a particular model of determination of bias, but his distinction may be applied to the larger literature of the topic. It is crucial to realize when critics are referring to cultural loading as evidence of cultural bias, as well as to realize that such "armchair" analysis is notoriously likely to be proven empirically wrong (Clarizio, 1978, 1979; Flaughner, 1978; Reschly, 1981).

Bias must be discussed in a context of validity. In current validity literature, the classical distinctions among empirical, construct and content validity appear to be disappearing, with all validity measures assumed under the umbrella of construct validity, and never being considered in isolation (Cole, 1981; Haney, 1981). However, the literature on test bias goes back more than fifty years and has dealt conceptually with validity in the following ways:

Construct validity. Construct validity concerns how well the score obtained from the test may be said to measure a theoretical construct or trait. Construct validity is always inferred, typically from such evidence as age differentiation, correlation with other tests purporting to measure the same construct, factor analysis, internal consistency, effect of experimental variables on test scores, and convergent and discriminant validation (i.e., correlating with measures with which it should theoretically correlate and not

correlating with measures with which it should not be related)
(Anastasi, 1968).

Content validity. Content validity concerns how well items of a test correspond to the behavior which the test attempts to assess. In the case of achievement tests, this may be known as "curricular" validity, that is, the degree to which the test measures the things that are intentionally taught (Haney, 1981).

Criterion-related validity. The criterion-related validities are concurrent validity--how well the test score correlates with some independent assessment of a criterion, and predictive validity--how well the test score correlates with some future assessment of a criterion.

Face validity. This is a subjective determination of the degree to which the items of the test would appear to reasonably relate to the concept the test is designed to assess.

Along with the differing concepts of validity there are differing concepts of bias. The following definitions of bias are typical of those found in the literature:

"Cleary et al." definition. "A test is considered fair for a particular use if the inference drawn from the test score is made with the smallest feasible random error and if there is no constant error in the inference as a function of membership in a particular group" (Cleary, 1968; Cleary, Humphries, Kendrick & Wesman, 1975). This is apparently the definition used by the court in Griggs v. Duke Power

Co. (1971), which required selection devices on which minorities average below the white majority to be shown to be related to job performance (Schmidt & Hunter, 1974). Four statistical tests must be passed to be fair under this definition: (1) equal variances, (2) equal, nonzero correlation coefficients between test score and criterion, (3) parallel regression lines, and (4) equal intercepts for the regression lines (Mercer, 1979a).

"Thorndike" definition. A test is fair only if for any criterion level the test selects the same proportion of the subgroup's applicants as would be selected on the basis of actual criterion performance. In use then, the percentage of the subgroup selected is the criterion for judging fairness rather than demanding equal mean scores, and the use of separate regression lines is acceptable (Thorndike, 1971). In practice, this definition may lead to "reverse" discrimination, since it is fair for groups but likely unfair for many majority-group individuals (Schmidt & Hunter, 1974).

Internal test structure bias. This type of bias evaluation considers whether scores behave similarly and in expected ways for all groups, for example, in prediction x job performance (Cole, 1981).

"Equal means" definition. According to this definition, test results must show equal group means to avoid bias. Mercer (1979a) posits the use of the Cleary et al. (1975) definition. However, in practice, if application of the test results in unequal means (Darlington, 1971) then she apparently considers the test biased (Mercer, 1973, p. 120ff).

"Larry P. v. Riles" (1979) definition. Citing Section 504 of the Rehabilitation Act of 1973 and the Education of All Handicapped Children Act of 1975 (which regulations require that tests used for evaluation be "validated for the specific purpose for which they are used" [34 C.F.R. §104.35; 34 C.F.R. §300.532]), the court set the following standards for intelligence scales used for assessment of black students: (1) same standard of scores for different groups, (2) approximately equal means for different groups, and (3) scores correlated with relevant criterion measures, i.e., classroom performance (Bersoff, 1981).

In addition to these broad definitions of bias, there are also definitions of bias which specifically relate to individual test items:

"Cleary" definition of item bias. An item is biased for a group for which the group average score on the item is different than the average score of the majority group by an amount in excess of that expected from the total score difference (Cleary & Helton, 1968).

Statistical item bias. These methods are designed to locate items which show differential functioning for different groups, and presumably therefore measure different things (Cole, 1981). These methods examine the function of individual items within the larger test and are useful but can only detect anomalous items, not pervasive bias.

Facial bias. When wording or item format would reasonably appear to disfavor some group, it may be rejected as biased (Cole, 1981).

Construct validity-based judgment of bias. This type of item bias involves "judgments involving characteristics of item content that are irrelevant to the intended construct, but that might elicit different responses to items from different groups of subjects" (Cole, 1981).

How the Various Definitions Interrelate

The effect of the working definition of bias may have a profound effect on the testee's chance of selection and the test user's overall employee performance rate.

Individuals selected by the Cleary et al. (1975) definition will on the average perform better than those selected by the Thorndike definition. The Thorndike definition leads to increased failures of selectees and results in discrimination against more qualified majority group testees. Use of the Thorndike definition leads to the potential for political abuse via possible identification of a multitude of possible subgroups and ultimately an extremely low validity of the test as a predictor. On the other hand, use of the Cleary definition will in some cases result in an almost zero probability of selection of minority members, especially when the predictive validity is relatively low and the majority group has a higher mean score. That is, when the majority group has a higher criterion performance, the minority is disadvantaged in inverse proportion to the validity of the test. The Thorndike and Cleary definitions function with equivalent

results only when there are equal criterion performance means or where there is a perfect correlation between the test and the criterion (Schmidt & Hunter, 1974).

In a scathing rejoinder to the Cleary et al. position, Jackson (1975) points out that using such a definition would also result in an ever smaller number of minority applicants being chosen as the validity of the predictor rises. He terms the report as "blatantly racist," and employing "a definition of 'fairness' which is intrinsically unfair" because the Cleary et al. report assumes that tests can in some way measure intelligence rather than the individual's cultural training to respond to items in a manner similar to that of the dominant white Anglo population. Bernal (1975) also challenged the Cleary et al. (1975) article as being an "apologia" for current testing practices.

The courts have tended to use the extremely simplistic notion of test bias that unequal means indicates bias, of which the Larry P. decision is fairly typical (Reschly, 1981). However, Lora v. New York Board of Education (1978) specifically disallowed such an interpretation, stating that "mere disproportionate impact along racial lines will not, in the absence of a finding of intent to discriminate, support a finding of a violation of equal protection guarantees." It has been almost universally agreed by psychologists that an equal means criterion is inappropriate because of the naivete of assuming that all groups will be equal on all criteria (e.g., Clarizio, 1978, 1979; Cleary et al., 1975; Flaughner, 1978; Bersoff, 1981; Reschly, 1981).

Evaluation of Current Tests According to Definitions of Bias

Using the Cleary et al. definition of test bias, the literature is almost uniform in its findings of no substantiated bias against blacks. Black score means are uniformly lower on standard "intelligence" tests, Piagetian tests, culture fair and nonverbal ability tests. In a survey of the literature, Kaufman (1979) states that the WISC-R measures the same basic abilities for minorities and whites (construct validity) and is equally effective in predicting school success (predictive validity). Other currently used intellectual measures generally pass technical criteria for avoidance of bias against minority groups (Reschly, 1981).

There is some evidence of bias against Mexican-Americans, who obtain a lower mean on the culturally loaded Peabody Picture Vocabulary Test than on the Raven Progressive Matrices Test (Clarizio, 1978, 1979).

Studies show that predictive validity is generally colorblind. Where differences exist, minority test taker's criterion scores are over-predicted (Cleary et al., 1975; Cole, 1981). The Riverside data (Clarizio, 1979) show equal predictability for blacks, Mexican-Americans and Anglos. There seems to be little evidence for charges of predictive bias against an academic criterion. Flaughner (1978) reports that currently used intellectual tests are free from bias with respect to both single-group validity (i.e., validity coefficient significantly above zero for one group but not another) and differential validity (i.e., significant difference between two groups'

validity coefficients). A general assertion may be made that tests are not psychometrically biased (Gordon & Terrell, 1981).

Flaugher (1978) points out that the Cleary et al. model of fairness is the only model listed that does not involve the use of a double standard for minority groups. Reacting to this violation of the principle of equal opportunity, McNemar (1975) suggested that, rather than lowering the requirements for minority groups, higher requirements should be imposed to eliminate the overprediction. Obviously, this is an area where technical expertise conflicts with social policy with emotional consequences (Gordon & Terrell, 1981).

The two widely used methods for examining statistical item bias are an Anova Item x Group interaction method and the item-response theory approaches. A prominent Anova interaction would indicate that items were operating differently for different groups, ergo, not meaning the same thing. The item-response approaches examine group differences by relating overall ability level to probability of correct response to individual items. However, differences found using statistical item bias methods have rarely been interpretable by study authors or expert panels. Also there is only moderate agreement in identification of "biased" items using the different methods, so evidence of bias using this approach tends to be at best obscure (Cole, 1981).

Using the facial bias method, Wechsler's "What is the thing to do if a boy (girl) much smaller than yourself starts to fight with you?" item is notorious for its supposed discrimination against non-Anglos. Empirically, however, Clarizio (1978, 1979) reports that the item is

actually psychometrically easier for blacks than whites. Tests may also have some subjectively identified items removed in an attempt to reduce potential overall bias. However the typical finding (Flaughner, 1978) is that removal of suspect items does not improve the performance of minority groups over the original version. Perhaps partially because of the high rate of disagreement among "experts" regarding what items are identified as biased, there seems to be no replacement for gathering empirical data to identify biased items. Flaughner also reports that this vein of study has also been generally disappointing. His general assessment of item-specific "fixes" of tests is that this is just not a promising option for overcoming differences in group means: "Something more is involved beyond simply asking the wrong questions."

Flaughner (1978) reports that where bias exists, it typically is in terms of user overinterpretation, in that more meaning is assigned to the test scores than is substantiated. For example, a test of verbal ability might be interpreted as showing overall poor cognitive skills.

When the relationship between predictor and criterion measure is the focus of interest, one must not neglect the possibility that the criterion itself is possibly biased. If the criterion measure has low reliability such that differences between group means are minimized, then one may infer bias in the predictor (because of larger differences than on the criterion) unless one determines if the differences would be more similar if both measures had equal reliability (Flaughner, 1978).

Mercer (1973) considered the typical criterion measures used to validate intelligence assessments (psychometric achievement) to be as biased as the intelligence tests, and proposed that a more relevant criterion would be teacher ratings. Clarizio (1978, 1979) criticized this choice on the basis that teacher standards are imprecise and that teachers' strong child-centered orientation makes them veer away from negative ratings. He also cites four studies which reported high correlations (from .32 to .83) between IQ and teacher ratings. Mercer reported finding only one study with correlations between .12 and .25.

Several psychologists have charged that the testing situation itself is biased against the minority examinee. For example, Jackson (1975) states that "psychological testing historically has been a quasi-scientific tool in the perpetuation of racism on all levels of social or economic intercourse." American psychometrics are seen as inherently racist because of the founding influence of Galton and Burt, and "IQ tests are no more and no less than samples of material presented in the racist education curricula The tests are more biased because, though biased as the racist interviewer, they purport to be 'objective.'"

With respect to examiner effects on test outcome, Clarizio (1979) reports that the literature is inconclusive, but the data usually indicates that blacks are not penalized by being examined by a white examiner. Also, dialectical differences between examinee and examiner show little if any impact on test outcome. The evidence is that the vast majority of inner city black youngsters can understand standard English as well as what might be their preferred dialect, perhaps

because of the influence of television. While the evidence is too skimpy to apply for all minorities, Clarizio (1979) reports that "general research findings seriously challenge the assumption that blacks in the United States do not share a common background of verbal and cognitive experiences which is sampled by standardized individual intelligence tests."

Flaughter (1978) cites evidence that potential bias inherent in the examination process is readily overcome with routine professional practices (rapport establishing techniques, etc.), plus that attempts to substantiate such a bias are so unfruitful as to make this line of study unpromising. On the other hand, he questions whether the high level of news coverage awarded to publication of test results do not possibly bias the educational attempts of educators who are working to raise achievement scores in some inner city settings, by engendering a feeling of futility in the students.

While acknowledging that test scores as predictors are not demonstrably psychometrically biased (Cleary et al. definition), a novel charge may be made that current tests' high predictive validity tend to make selection practices choose fewer minority applicants than would be the case if the tests were unavailable (Jackson, 1975). Some critics have charged that tests tend to become the sole criterion for access to education and employment because of their high validity, and in so doing minimize the likelihood of selection of somewhat "higher risk" examinees (Gordon & Terrell, 1981).

Infrequently mentioned in the literature, probably because of its obvious impact on interpretation, is the issue of using standardized

assessment devices on students with organic handicaps. Only the most careless examiner would attempt a nonqualified interpretation from test scores obtained by a subject with a sensory impairment.

Obviously, the interpreter can make allowances, but there is no assurance that obtained scores will have the same meaning as scores of non-handicapped subjects (Green, 1981).

As Cole (1981) has pointed out, advocates of neither side of the testing bias controversy are extremely well supported by the data in the literature. Test defenders find "frequent although inconsistent differences" in the way in which tests function for different minority groups and anomalous items. Reformers find lack of clear evidence of bias, but enough incongruities to make abandonment of their mission unlikely. Reformers can take credit, however, for raising the awareness of test constructors to issues of fairness and for the cleaning up of a great deal of facial bias in published standardized tests.

Ultimately, the controversy of test bias is not going to be settled through scientific inquiry (Gordon & Terrell, 1981). The rationale of ultimate appeal for opponents of current testing practices is that the "yardsticks" are unfair in their conception and application because of their practice in the sociopolitical context of minority groups being overrepresented in the disadvantaged classes. From such a viewpoint, the purpose of testing is itself discriminatory because it restricts opportunity to achieve "against the odds." For example, in an article in a National Educational Association publication, public advocate Ralph Nader charged the Educational Testing Service with bringing to an end "the Horatio Alger period in America."

The same article quoted Chuck Stone, former director of minority affairs at ETS, describing his former company as "unopen, unyielding, unaccountable, unconstitutional, unequal, unresponsive and unreceptive" ("Teachers and citizens protest the testing ripoff," 1980).

Proponents argue that the "yardsticks" are proven to be "scientifically" (i.e, psychometrically) valid and thus beyond any rational criticism as being biased. This viewpoint is of course from a traditional psychometric context with little or no concern for political and social issues.

It then becomes essential to distinguish between the questions "Does the test accurately test what it purports to test without bias?" and "Is it in the society's interest to use the test for the proposed purpose?"

Cultural Pluralism and Mental Testing

In reference to standardized group tests, president of ETS William Turnbull says that testing first became significant as an effective method of discovering and encouraging talent. Only later, as competition for admission to college mounted, were tests seen in a new light—as a possible barrier to higher education. In the interim, some people had imputed a near infallibility to test scores so that realistic ideas about testing suffered more from the hands of its proponents than from its critics. The ensuing "discovery" that the accuracy of tests is not perfect has obscured the fact that they are more reliable than most of the alternatives (Educational Testing Service, 1979). Turnbull asserts elsewhere that the real issue is that scores on tests tell you nothing about the odds overcome to make

a given score, nothing about the testee's energy and determination (Samuda, 1975).

Opponents to current testing practices sometimes adopt the view that tests serve to maintain the "myth" of meritocracy in society (Aronowitz, 1974; Evans, 1974; Kagan, 1974; Purvin, 1973). (For the archetypal meritocratic statement, see Herrnstein [1971].) These social critics maintain that the powerful classes only create an illusion of upward mobility, persuading "upward risers" to substitute better paying jobs in exchange for meaningful impact on the conduct of society. Aronowitz (1974) states that the concept of intelligence and the meritocracy myth are predicated upon each other. In so doing he condemns the "liberal humanists" (the environmentalists in the heredity-environment debates) for accepting the prevailing social division of labor implicitly, and in so doing, condemning those they would defend:

The IQ test symbolizes America's interpretation of equality to mean equality of opportunity In this context environmentalists implicitly assume the standards set by the standardized IQ measure—namely, access to jobs as reward, not access to more powerful participation in determining how the pie is to be divided.

Turnbull (Samuda, 1975) is probably correct when he states that intelligence testing has captured increased controversy because the definition of intelligence is not standardized, the intellectual measures have become of increasing practical importance in competitive situations, minorities have experienced increased pride and self-awareness and refuse to be evaluated by what is seen to be "white Anglo-Saxon Protestant" criteria, and the resurgence of the heritability issue.

Important Assumptions Underlying Intelligence Testing

Intelligence testing assumes that what the individual has learned in the past is a valid predictor of what he can learn in the future. The tests further assume that all persons tested have had equal opportunity to learn the materials to be found in the test samples. Although all examinees have had equal opportunity to learn, people master the materials to varying degrees. The obtained test score differences reflect then the differences in the underlying intellectual ability (Rapier, 1962).

Mercer (1978) offers the following paradigm for the circumstances in which an interpretation may be made about a subject's "intelligence" based on his score on an intelligence test:

IF two persons have had equal opportunities to learn the materials in a test, and

IF they have been equally motivated and reinforced for learning the materials in a test, and

IF they are equally familiar with tests and test-taking situations, and

IF they are equally free of anxieties, fears, or emotional disturbances that might interfere with learning or with test performance, and

IF they are equally free of physical, sensory or motor disabilities that might interfere with learning or with test performance,

THEN significant differences in their test performance may be interpreted as reflecting differences in their learning potential, i.e., their intelligence. (p. 53)

Standardization

Until the mid-1970's, the most used IQ tests were the Stanford-Binet and the WISC, neither of which used any black children in standardization. During the standardization of these two tests,

nonwhite and foreign-born children were excluded according to the rationale that their inclusion would result in unduly low norms which would not be predictive of the majority population (Evans, 1974). Yet the critics point out that these same tests have been routinely used to label and assign minority students to EMR classes.

The Stanford-Binet was deliberately constructed to eliminate sex differences because the sexes were assumed to have equal intelligence (Terman & Merrill, 1973). Critics have been quick to point out that ethnic differences could have as easily been eliminated, if the test constructors had not assumed the intellectual superiority of the white population (Evans, 1971).

During the restandardization of the WISC (the WISC-R), a proportional sample of blacks was used but are massed into the total (Wechsler, 1974). Group tests such as the Otis-Lennon and Lorge-Thorndike are standardized on larger samples (for example, the sample size of the Otis-Lennon is 200,000). Yet even these tests' normative samples assume that ethnic groups are uniformly distributed across the social classes. Thus these scales may not be seen as adequately representative of the nonwhite population (Samuda, 1975).

Children from lower class homes tend to be from quite diverse environments because of inconsistencies in school experiences, more frequent relocations of the family, less stable family composition, and fewer "educational" materials in the home. These factors tend to make the performance of lower-SES children more variable and less predictable from the basis of test scores (Rapier, 1968).

Testing critics flatly reject the notion that the lower group means of some ethnic minorities reflect lower mental abilities. Intelligence tests in current use were largely developed and standardized on white middle class children, and that genesis is seen as the source of the bias. The test development assumes that white standards, values, attitudes, beliefs, experiences and knowledge are the only "correct" ones, thereby stripping ethnic and economic minorities of their cultural distinctiveness and labeling their distinctiveness as "defective" (Samuda, 1975). Larry P. v. Riles et al. (1979) and Hobson v. Hansen (1967) both involved judicial decision of "irreparable harm and injury" being made to students wrongfully labeled and placed in special education classes on the basis of inappropriate evaluation.

A position statement issued by the Council for the Society for the Psychological Study of Social Issues points out that black and white test score differences diminish as one controls for "cultural" and educational background—"the more comparable the background, the less the difference (Council . . . , 1969). Low scores on intelligence tests and poor performance in the schools may be for some students a positive test result of having learned well from school and family the appropriate behavior for the role allotted them by society. Poor performance on an intelligence test is seen for poor people evidence of "successfully socialized personalities" (Gartner, Greer & Riessman, 1974).

The President's Committee on Mental Retardation in a 1970 memorandum (Gerry, 1973) viewed overrepresentation of minority groups

in special classes as due to two major problems: (1) a failure to understand, value and utilize cultural and linguistic backgrounds of minorities, and (2) a "conscious or unconscious effort to retain minority groups in subordinate status."

Those who see in present practices of intelligence testing a strategem by the power structure to maintain control over the less powerful have their spokespersons. Gartner et al. (1974) wrote that

it should be no surprise that at a time when the social policy of dominant groups is to curtail mobility, to thwart change, to block redistribution of power and wealth, the IQ test is once again hailed as a valid and objective tool. But IQ as a synonym for cognition is as inaccurate as is the faith in the relation of school, merit, and opportunity itself. As for those who seek to assure opportunity, they must first demythologize the IQ test. (p. 6)

Among those who see current testing practices as a tool of class suppression, Garcia (1981) rejects the test-bias literature's apologetics as a reification and justification for the upper classes keeping their socioeconomic standing. He specifically rejects the arguments built around minorities' scoring uniformly lower on both verbal and nonverbal tests, that minorities have the same regression lines and are predicted on criterion outcomes equally accurately. (The same end could be achieved, Garcia points out, by tests biased in favor of minorities.) The argument that tests play a critical role by measuring the student's adaptation to the environment in which the child must function if he is to succeed is particularly insulting to minorities: the ethnic groups do have varying environments which are viable alternatives to the majority culture.

A host of explanations are made for minorities doing less well on intellectual ability tests. Among the most common is that intelligence tests do not measure "intelligence," but instead measure the compatibility of the testee's language and culture to that presumed by the test, the skill of the tester in using language and interpreting responses in a manner appropriate to the circumstances of the testee, the familiarity of the testee to behaviors valued by the test (for example, quick responses and Aristotelian logic), the anxiety level of the testee, and the testee's motivation to succeed on the test. Furthermore, it is charged, the test administrators often expect failure and this "halo" effect affects the scoring of the protocol. The overreferral of minorities for evaluation for possible special education placement reinforces the examiners' expectations of failure. Also, the poor communication between the schools and parents leads to overlooking of other possible problems such as vision, hearing or emotional problems (Gerry, 1973).

Samuda (1975) cites three core issues concerning the testing of minority group members: (1) the lack of reliable differentiation in the range of minority group scores, which tend to cluster at the lower end of the total range, (2) the lack of predictive validity when scores are compared to standardization samples of a different socio-cultural background, and (3) an emphasis on adequate understanding of the sociocultural background of the group being tested in order to make a valid interpretation of scores.

Fishman, Deutsch, Kogan, North and Whiteman (1964) report three main difficulties in using standardized tests with minority groups.

(1) The tests may not reliably differentiate in the score range where most of the group's members may score. This is because a smaller range of scores results in lower psychometric reliability. (2) Test scores may have different predictive validities for minorities. The test-related factors affecting the obtained scores (but not the criterion) include

test taking skills, anxiety, motivation, speed, understanding of test instructions, degree of item or format novelty, examiner-examinee rapport, and other general or specific abilities that underlie test performance but which are irrelevant to the criterion.

Factors associated with the complexity of the criterion measured include motivation, class behavior, personal appearance and study habits. Differential validity can also result from the nature of the intervening events and contingencies. The situational, motivational and maturational effects of the unreinforcing climate of the minority's cultural milieu may be seen as the cause for the progressive decline in achievement test scores of such children as a function of time.

Fishman et al.'s third difficulty in using standardized tests with minority students is that adequate interpretation of the tests depends highly on the interpreter's understanding of the social and cultural background of the testee's cultural group. Fishman et al. (1964) list possible sources of interpretation error to be deviation error, simple determinant error, incompleteness of data error and failure to consider the testee's chance for exposure to the test content. Deviation error is the inference of maladjustment from responses made by the examinee that would be deviant for the test's normative group but are not deviant with respect to the examinee's

group. The simple determinant error arises from construct and content validity, and is viewing the test as reflecting some pure construct without respect to the population being measured or the conditions of measurement. Failure to measure intellectual factors important to success is the content validity incompleteness of coverage error.

Fishman et al. report that lower class students tend to have certain characteristics which attenuate their test scores and the validity of the scores, including observations that lower class children are less verbal, more fearful of strangers, have lower self-confidence, attain lower academic achievement and have less intellectual competitiveness. The lower class child is likely to have greater "irritability" and to be less conforming to middle class behavioral and conduct norms. He is also more likely to be bilingual, have fewer recreational outlets, have less knowledge of the world outside the immediate neighborhood and more likely to attend substandard schools. This latter finding may be a factor in the reported findings (e.g., Deutsch & Brown, 1964) that blacks born in the North obtain on the average higher IQ scores than blacks born in the South, and that longer residence in the North yields higher IQs.

Gay and Abraham (1973) detail relevant differences between the black and Anglo cultures including attitudes toward whites, socialization patterns, cultural clash, communication styles and attitudes toward interactional patterns. In a statement highly relevant to the interest of this paper, they state that

Admittedly, suspiciousness of middle-class, mainstream whites may be common among all low socioeconomic groups, irrespective of their ethnicity Much of what we have said thus far could apply to a wide variety of post-

agrarian households, including numerous 'white ethnics' like Appalachians, Italian-Americans, Irish-Americans, etc.

There are cultural factors which are commonly cited reasons for why poor and lower SES homes produce low-IQ children. Among these are fearfulness of the testing process, expectation of failure, insensitivity to speed requirements, poor test taking skills and lack of familiarity with the test item problem content (Budoff, 1975).

Riessman (1974) posits that the assumptions upon which intelligence tests are constructed build in a bias against members of "deprived" groups. Intelligence tests are generally designed to be brief. The style of the deprived child is generally to be slow and cautious, taking time to become involved in a problem. This emphasis on speed works particularly against rural children. Second, IQ tests are scored against lists of correct answers, not correct reasoning used to generate answers. Research shows deprived children to be poor judges of distractors, presumably because of limited test taking skills. Third, test items not showing improvement with age are excluded. The deprived child's IQ typically drops with age, perhaps because of a lack of exposure to vocabulary and experience of the larger culture. If IQ tasks were not dependent upon these experiences, the deprived child might do better. A fourth biasing factor is the intelligence tests' emphasis on academics. Deprived children may have more practical experience with practical problems of the real world than the middle class book-oriented youngster. Yet schools do not have broad enough goals to use those skills.

Haggard (1954) found four critical weaknesses in intelligence tests:

Among other things, (a) they have measured only a very narrow range of mental abilities, namely, those related to verbal or academic success, and have ignored other abilities and problem-solving skills which are perhaps more important for adjustment and success—even in middle-class society; (b) they have failed to provide measures of the wide variety of qualitative differences in the modes or processes of solving mental problems; (c) they have ignored the influences of differences in cultural training and socialization on the repertoire of experience and the attitude, motivation, and personality patterns of subgroups in our society, and the effects of such factors on mental test performance; and (d) they have considered mental functioning in isolation, thus, ignoring the interdependence of the individual's motivational and personality structure on the characteristics of his mental functioning, as seen, for example, in the differences between rote learning and the ability to use previous experiences creatively in new contexts. (pp. 180-181)

In his 1954 study, Haggard reasoned that poor children are hampered by lack of knowledge of how to take tests, were poorly motivated and were distant from and afraid of the examiner. Haggard's design included giving both poor and middle class children three one-hour training periods in taking IQ tests, giving rewards for doing well, and training examiners in techniques to enhance rapport. The IQ's of the poor children rose 15 to 20 points on the average, while the middle class student's gains were much smaller, presumably because they were already performing nearer capacity. He notes that lower class students "will try to do well in tasks involving their particular value systems, but superior performance on intelligence tests generally is not one of them."

Jensen (1968) posits that intelligence tests do a fairly reasonable job of reflecting actual differences in ability between groups, and suggests that the environment offers an explanation for a significant part of the lower verbal ability scores of lower socioeconomic groups. The literature reports that lower SES families tend to engage

in less verbal play and verbal interaction, and have lower rates of reinforcing the early vocalization of infants. There is relatively little "labeling" or naming play. There is less concerted "shaping" of speech by lower class parents. The larger family size which typifies many minority groups results in more children sharing the attention of adult speech models and greater time using peers as verbal models. Many minority parents are less likely to respond to questioning asked by their children. Lower SES speech patterns bear less resemblance to written language than do the normative speech patterns, so the minority members reading and written language can be quite different from his verbal language. Minority oral language is reported to allow for fewer nuances of speech and less abstraction than normative speech.

Rohwer (1968) reports that "retarded" learners use different learning strategies than their nonlabeled classmates. He relates the use of the mnemonic strategies of naming, elaboration and use of pictures, words and objects to early environments and thus to differences between and among SES groups.

Samuda (1975) suggests that minority students may perform less well because of factors extraneous to the test, such as testee motivational level, anxiety, suppression of aggressive tendencies and perceptions of hostility in the testing environment. A basic argument may be made that present testing practices allow the discovery of talents and ability only of members of the higher SES groups, and miss the abilities of persons from lower SES groups. Aronwitz (1974) states that

a good case could be made that IQ tests really inversely measure authentic problem-solving intelligence and reasoning; that doing well on these tests implies a trained incapacity to think in depth, to think innovatively, to be creative.

Particular test items are frequently used to demonstrate bias.

Jerome Kagan (cited in Lawler, 1978) cites the question of "What should you do if you were sent to buy a loaf of bread and the grocer said he didn't have any more?" To get credit, the child must say that he would go to another store; a response of "Go home" earns no credit. The situation where the first response is the "intelligent" response, however, excludes those situations where no other store exists within safe walking distance, where grocery purchases are dependent upon the operator extending credit, or where the student has been explicitly ordered to go home in such a case. As Kagan notes, "the child is asked how a piano and violin are alike, not how a tortilla and frijole are similar."

In response to this line of reasoning, Jensen (1970) points out that despite obvious cultural loading of some test items based on face validity, it has been found that the "obviously" culturally-biased items are not those giving lower SES children the most trouble. They often do worse or no better on tests with less cultural loading such as Raven Matrices. He cites Eells, Davis, Havighurst, Herrick and Tyler (1951) as finding that the one discriminating factor in culturally discriminatory items was the degree of abstractness of the test items.

Consequences of Inappropriate Testing

Fifteen years ago, Rosenthal and Jacobson (1968) published their well known and highly criticized study (Mason, 1975; Guskin, 1978),

which purportedly found that leading teachers to believe that certain students would show unusual intellectual development resulted in the prediction coming to pass. This book, Pygmalion in the Classroom, helped institutionalize the phrase "self-fulfilling prophesy," which has come to mean that individuals can be predisposed for certain expectancy effects and in some fashion unintentionally bring the expected outcome to attainment.

Teacher bias and expectancy effects figure prominently in the sociological labeling literature dealing with the phenomenon of overrepresentation of minorities in classes for the educationally handicapped. The following is a statement in relatively pure form of the perceived scenario:

A student is tested; the test may be, for a variety of reasons, inaccurate, and the test results often characterize the student as being below normal in intelligence. This characterization functions as a label which goes to the teacher and prompts attitudes and treatment that may be inappropriate and false and even worse, are subtly (or otherwise) transmitted to the student. The student, in turn, tends to reflect behavior that fulfills the teacher's expectations. In short, if the psychologist tells the teacher that Johnny is below average and if the teacher treats him as such, Johnny will respond accordingly. . . . This situation is still being repeated countless times in schools throughout the nation: the issue is what to do about it. (Meeker & Meeker, 1973)

Oakland (1973) charges that current tests imperfectly predict behavior, rigidly shape curricula and restrict educational change, foster views of innate and fixed abilities and foster undesirable biases and expectations. He finds the relevance of norm-referenced tests to instruction rather tenuous and reflective of only middle class values and attitudes, which aids in the maintenance of racial bias and stratification. Since currently available tests do not tap nonmajority

culture values, attitudes, and linguistic/cognitive abilities, the test administrator is never able to know whether the child with tested "low" linguistic skills is deprived linguistically or whether he uses a different linguistic system (Pyle, 1979).

One of the worst possible consequences of inappropriately used intellectual testing is the possibility of legitimizing poor student progress and excusing the school from responsibility to try to overcome the difficulty (Pyle, 1979; Laosa, 1976). Brody and Brody (1976) express a warning that both professional psychologists and educators are capable of creating great harm to students because of lack of adequate knowledge of the limitations of tests and the proper inferences that may be made from them. They further charge that the dangers from use of the tests may not be outweighed by the advantages, since intelligence tests in current use offer in their opinion little or no diagnostic use for providing insights into specific intellectual processes and ultimately legitimize inefficient and caste-like grouping of students on the basis of "ability" (Findley, 1975).

As the test interpreter leaves direct correlational inferences and enters the interpretive arena of "clinical inferences," an ever-mounting potential for bias arises. Inferences and implications based on test observation, item responses and subtest score patterns are "rarely, if ever, supported by appeal to relevant empirical data" (Brody & Brody, 1976). Matarazzo (1972) states,

Alas, hundreds upon hundreds of studies on the use of profile, pattern or scatter analysis with the Wechsler scales conducted between 1940 and 1970 failed to produce reliable evidence that such a search would be fruitful.

Furthermore, the research on subtest interpretation tends to generally have independent and dependent variables reversed. Instead of knowing what percentage of a deviant subgroup possesses characteristic "x" (which is what is reported), we should like to know what percentage of the population possessing characteristic "x" belongs to the deviant class. Applying clinical hypotheses to minority classes would appear to multiply the chance for error.

Labeling

In the Riverside study of the identification process of mental retardation, Mercer (1973) found that most persons labeled retarded were identified as such by the school system, and that those so labeled by the schools were very often not so identified outside the educational setting. She also found that poor and minority students stood a much larger chance of being labeled retarded, and concluded that this disproportionate identification was due largely to use of standardized intelligence tests. This researcher and her study played a key role in the Larry P. v. Riles (1979) decision, which banned the use of norm referenced intelligence measures with black students.

Several general conclusions may be drawn about labeling (Guskin, 1978). There is concern about negative associations to the label "mental retardation." There is concern that children labeled retarded will be seen less favorably, and that these children will then be treated in a less pleasant or appropriate way. Finally this behavior from the environment may have negative consequences for the achievement and personal adjustment of the child labeled retarded. In a comprehensive review of the literature of teacher bias and expectancy

effects, Guskin (1978) concluded that labeling per se seemed to have a minor effect on how others dealt with the individual, except in those situations where the label was all that was known of the person.

People tend to use behavioral cues as a more important determinant of how they will relate to other people than the label "retarded."

However, this is not to suggest that mislabeling is in any way to be taken lightly. It is just not the devastating influence it is assumed by some to be.

Sociology, specifically sociology of deviance, provides a distinctive theoretical formulation and study of labeling. Two main branches of labeling theory (Guskin, 1978), attribution theory and the labeling theory of deviance, are helpful. Attribution theory examines how laypersons perceive the cause of the observed deviant behavior. For example, deviant behavior may be assigned to be caused by ability, effort or luck. The labeling theory of deviance is expounded by Mercer (1973). In this model primary deviance (observable behaviors which result in society reacting to the individual as someone deviant) are less important than secondary deviance, the new set of expectations and behaviors which are maintained once the individual has been defined as deviant. According to this theory, the person is not "mentally retarded" until labeled as such by himself and/or others. Labeling may be made formally or by differential treatment, and results from failing to meet certain competence standards expected of normals. Thus another society might not label a person's behavior as deviant, and the person would therefore not become retarded. Once the person is labeled retarded, however, he enters into a "career" of

retardate and must behave in accordance with the social role of a retardate. He may adopt further deviant behaviors which are a function of role and expectations.

Test results can reinforce social roles by placing students into poor performer categories, keeping them there, infusing an "inferior" self-consciousness, and maintaining a class dependency (Samuda, 1975).

Guskin (1978) offered this overview of the current status of labeling theory:

The rash of interest in labeling seems to have developed out of a social orientation which involves identification with those who have been rejected by the larger society: minority groups, the poor the criminal, the mentally ill, and now, the mentally retarded. This identification with the downtrodden has been accomplished by the attribution of their difficulties to society and its institutions. Whereas earlier writings had sought the determinants of these problems within the deviant individuals and later writings had attributed the causes to a generally deprived social environment, current tendencies hold society directly responsible for deviance through the very agencies developed to treat the problems. These are seen as creating deviance through labeling and their reactions to the labels. Thus intelligence tests, special education classes, and residential institutions for the retarded have become defined as causes of, rather than solutions to, the problems of mental retardation. This change in social perspective is reflected in sociological and social psychological theory and research as well as in political decisions and applied writings relevant to the treatment of the mentally retarded. (p. 128)

Minority students, males and poor whites are overrepresented in special education classes (and thus necessarily labeled) by factors as large as 3 to 4 times predictable incidence. Reschly (1981) is representative of those who see the basic problem with overrepresentation of minority group members in classes for handicapped students not so much with IQ tests as in the poor quality of the available special education programming and the labeling effect. However, he has harsh

criticism for the use of IQ testing, that is, the psychoeducational evaluation process as currently practiced: "humiliating and stigmatizing for the individual, racist with respect to groups, and most important, irrelevant and damaging to educational programming."

Reschly cites multiple incidents of abuse of the evaluation process. Bilingual students have been placed into EMR classes on the basis of verbal IQ scales administered in English. The scales used have been short form and group tests administered by poorly trained personnel. Sometimes parents have not been notified of either the referral, evaluation or placement of their children for special education services. And special education programs have been typified as having little academics, poor facilities and poorly trained teachers.

The assumption of many court decisions has been that the intelligence test is the primary basis for classification of mental retardation (Mercer, 1973). The implicit assumption of Larry P. v. Riles (1979) is that persons associated with the conventional use of intelligence tests, and school systems with disproportionate numbers of minority students in special education programs, are proponents of genetic explanations of racial differences. In the Larry P. decision, Judge Peckham clearly assumed that non-identical distributions were the result of "bias." A chief expert witness for the plaintiffs, Mercer (1979a) writes that psychologists must either accept pluralistic norms or engage in discriminatory practices.

By assigning all the responsibility for disproportionate assignments to special programs upon intelligence testing, critics tend to

ignore the fact that referrals for evaluation are made on the basis of school failure (Reschly, 1981). Yet an important role to be played by tests is to minimize bias. Bernal (1976) stresses the importance of a non-biased assessment process because many referrals for special education evaluation result from the normal display of some cultural expression which is unfamiliar to the referring agent, or perhaps because some teachers may not approve of some pattern of activities common to some ethnic groups. Findley (1975) cites a study by Carrager which examined one school system's method of placing students into a high academic track with a procedure involving the use of test results. On the basis of testing alone, 77 percent of the high SES group qualified for the top track and 38 percent of the low SES group. When grades plus teacher and administrator judgment were added to the criteria, the low SES group members fared even more poorly. Eighty percent of the high group continued to qualify for the top track but only 50 percent of the low group were awarded the top placement.

Recommendations for Changes in Evaluation Process

Even a casual survey of our friends and acquaintances would lead us to reject the notion that one [individual] is generally superior in all ways to another. General intelligence is a trait we apply to strangers; our friends are differentially specialized creatures. Testing is also for others, for the young and the powerless. The mark of success is to escape from this irksome rite of passage. But reform of testing is possible. (Garcia, 1981)

Given that present evaluation practices are not those which Pangloss (of Candide fame) would describe as being the best of all possible in the best of all possible worlds, what can be done to maximize the potential to be found in the assessment process? Specifically, how may psychologists minimize Type I error, i.e.,

falsely identifying a subject as retarded, without grossly inflating Type II error, i.e., failing to identify the presence of retardation (Hays, 1973)?

Cronbach (1975) details the racism rampant in the American society after World War I, during the period of rapid rise of the testing movement. After much debate more characteristic of yellow journalism than professional writing, the debate featured such prominent persons as C. C. Brigham (who believed Southern and Eastern Europeans to be subnormal intellectually), Robert Yerkes (who felt, among other things, that not more than 10 percent of the population was capable of earning the B.A., and their quitting schooling before earning it proved as much), L. M. Terman (who engaged in rather unscholarly printed sarcasm), and writer Walter Lippmann (who wrote of the testers' "self-deception as a preliminary to public deception"). Finally in 1923, F. N. Freeman (cited in Cronbach, 1975) published a consensus statement endorsed by Brigham and Yerkes which more or less cleared the air and stated, among other things, that there was no logical way to judge the genotypic abilities of groups with dissimilar environmental backgrounds.

The SPSSI statement on race and intelligence (Council for the Society for the Psychological Study of Social Issues, 1969) has been referred to previously. It is worth noting that Cronbach (1975) finds complete harmony between the Brigham-Yerkes-Freeman consensus statement of 1923 and the SPSSI statement of 1969.

In a previous section, a passage was quoted from Mercer (1979a) in which she dictated the logic involved in inferring differences in

intelligence from differences in test scores. The logic involved equal opportunities to learn, equal motivation, and equal freedom from emotional distractions and organic disabilities as prerequisite assumptions. Of course, the problem in applying this logical model is that these assumptions are never met.

A variety of strategies have been tried or suggested to ameliorate the invalidating influences. These approaches range from ultimatums calling for an outright ban on intelligence testing to a resolute defiance somewhat short of denying that anything is amiss in the technology of intellectual assessment. Between the extremes lie such approaches as the extension of criteria in the definition of educable mental retardation, attempted development of culture-free or culture-fair tests, development of culture-specific tests, development of culturally-pluralistic norms or adjustment of normative expectations of minority group members, development of Piagetian assessment instruments, retraining of examiners and adaptation of the clinical testing interview, innovative analysis of current intelligence tests, use of gain or learning scores, development of new types of intellectual assessment tests and use of criterion-referenced tests.

Samuda (1975) identifies four principal trends in the development of alternatives to traditional measurement practices: changes in interpretation of scores of minority subjects derived from their performance on traditional norm-referenced tests; attempts to develop tests whose content would be, theoretically, equally fair or unfair to different cultural groups; attempts to relate scores derived from the performance of minority subjects with measures of the environmental

circumstances of the individual students; and attempts to develop tests in areas of skill mastery with a minimum level of performance or criterion rather than comparing the individual's performance with the average score of the standardization sample.

Bernal (1976) believes that assessment practices could be greatly improved by greater sensitivity to cultural diversity and community during all stages of assessment. Parents "should be considered as active agents during the referral and assessment phases and as partners in the intervention process; they need not be passive permission granters or potential legal liabilities."

Aliotti (1980) calls for redefining intelligence testing as a cognitive interview in which the psychologist obtains a behavior sampling of those skills especially valued by the society, and interpreted in light of the problem solving strategies used by the subject, his interaction with the examiner and all other information available to shed light upon how the subject functions as a cognitive being.

Isaacson and van Hartesveldt (1978) feel that psychologists are too quick to ascribe mental retardation to students with poor academic records, writing that "children with very low marks in reading may be quite advanced students of ghetto survival, a course that most members of the middle class would fail." Concluding that many minority special education students were "6-hour retardates," the President's Committee on Mental Retardation detailed an exhaustive assessment process for nondiscriminatory assessment. The May 25, 1970 memorandum prohibits

the use of cultural and linguistic differences of minority children as a means of segregating or denying such children

equal access to the full benefits of the educational system and reflects the belief that school districts have a constitutional and statutory obligation to administer their educational programs with sufficient flexibility to assure equal access of all children to the programs' full benefits As a result minority children are not penalized for cultural and linguistic differences, nor are they asked to bear the unfair burden of conforming to a school culture by the abandonment of their own. (Gerry, 1973)

The President's Committee recommended the following assessment process for use in school districts where discriminatory assessment practices had been found: Before assignment for evaluation, the school district gathers and analyzes adaptive behavior and socio-cultural background information regarding the student's out-of-school environment. This analysis includes adaptive behavior ("the degree to which the student is able to function and participate effectively as a responsible member of his family and community"), the student's incentive-motivational style ("those attributes of the child which characterize the manner in which he is most likely to be motivated to learn"), and his learning style (the type of learning activity most likely to result in acquisition of desired skills). In addition, at-home and in-community language skills, interpersonal skills and behavioral patterns were to be assessed.

Calls for Moratorium of Testing

In 1968, the Association of Black Psychologists called for a moratorium of testing black youngsters until appropriate black tests were developed (Oakland, 1973; Cleary et al., 1975). In 1970, an American Personnel and Guidance Association Senate resolution called for a moratorium of testing minority members with group intelligence tests. In February of 1972, participants at the National Education Association's "Council on Human Relations" recommended that standard-

ized achievement and IQ scores be removed from students' cumulative records and that the testing industry should be censured for "not responding to the needs of culturally different children." Concluding that "tests cause psychological damage to all children, and especially minority children," the group called for a 3- to 5-year moratorium on all group and individual intelligence testing, and abolition of the term "IQ" along with group tests of mental ability. These resolutions were adopted by NEA's Representative Assembly at its annual meeting in 1973 (Bosma, 1973). By 1980, NEA was distributing anti-testing lobbying kits (titled Standardized Testing: The Frustrating System that Traps Our Students) to its local association groups ("Teachers and citizens protest the testing ripoff," 1980).

In 1974 the National Association for the Advancement of Colored Persons called for a moratorium on standardized testing until standardized tests were developed which were "corrected for cultural bias and which fairly measure the amount of knowledge retained by students regardless of his or her individual background" (Samuda, 1975).

George Jackson, writing as Chair of the Association of Black Psychologists, urged imposition of a moratorium on testing with government intervention and legal sanctions for violation (Jackson, 1975). Brody and Brody (1976) summarized that "we can find few if any positive functions for routine intelligence testing in the public schools, and some possible negative functions to which tests lend themselves. And, as a result, we do not believe that tests should be routinely used in the public schools." Fine (1975) called for the

elimination of all IQ scores from cumulative record folders to thwart prejudging and concomitant bias effects on the part of teachers.

(After current IQ tests are re-evaluated, restandardized and their validity re-established, he was willing that surviving tests be used provided that the tests are labeled "CAUTION -- THIS I.Q. TEST SHOULD BE USED SPARINGLY AND AT NO RISK TO THE CHILD.")

The use of tests with black students has been banned in California, as a result of the Larry P. v. Riles (1979) decision. Attempts to ban tests have also been tried in Minneapolis, Philadelphia and New York City (Gordon & Terrell, 1981).

Aliotti (1980) questions that the result of the California ban has been an improvement in nondiscriminatory practices since the methods replacing the intelligence tests are less reliable, less valid and unproven with respect to nondiscrimination. He states that

currently California school psychologists are utilizing Piagetian assessments, adaptive behavior scales, covert IQ testing in which the data is not reported, and comprehensive test batteries of achievement and perceptual tests to circumvent the court decision. Paradoxically, some of these tests and procedures have not been validated for the purpose to which they are used, that is, to assess the intellectual functioning of children and adolescents.

The call for a moratorium is rejected by Humphries (1975) because of the proven interpretive value of test data, and the robustness of the interpretations despite the issues of race, ethnic origin and SES of the testee.

Culture-Fair Testing Movement

Culture-fair tests have been recommended by many psychologists and educators in years past as an answer to assessing individuals'

intellectual ability without unduly contaminating the measurement with culturally biased factors. In practice culture-fair tests deemphasize factors which mitigate against members of lower socioeconomic classes such as speed, specific item content and heavy verbal loading. Culture-fair tests tend to be primarily nonverbal tests administered without use of written or spoken language, although directions may be given orally or through the use of pantomime. Materials are visual and items are selected to be equally common or uncommon to all groups. Speed is generally neither penalized or rewarded.

There have been many attempts at the development of culture-fair tests. The Army Beta test of 1917 was a nonverbal group test designed to not discriminate against nonreaders of English. Pintner and Paterson introduced the first performance tests for clinical use in the U. S. in 1917, the test using form boards and wooden puzzle boards. The Porteus Maze test appeared in 1915, and the Kohs Block Design test in 1920. Other attempts at culturally-nondiscriminatory tests include the Cattell Culture-Free Intelligence Test of 1944 and the IPAT Culture-Fair Intelligence Test of 1949 and 1963, a revised Cattell test which was based on Cattell's theory of fluid intelligence (Cattell, 1968). The so-called "Chicago group" (Samuda, 1975) developed culture-free tests from the theoretical position that removal of cultural influences would result in no significant differences in average intellectual ability despite socioeconomic distributions. Native intelligence was seen as lying "buried in pure form deep in the individual and [needing] only to be uncovered by ingenious mining methods" (Wesman, 1968). When "culture-free" tests proved impossible

to make, attempts were made to develop "culture-fair" tests, which did not try to exclude all cultural influences but rather to include only cultural components common to different cultures. One Chicago group product, the Davis-Eells Games Test of 1953, was withdrawn from the market because the lower socioeconomic students performed as poorly on it as on conventional tests, and because it proved to have low validity as an academic predictor (Samuda, 1975). Raven's Progressive Matrices of 1938 and 1947 were developed to be an uncontaminated measure of g. Research showed the Leiter International Performance Scale to be neither culture-free nor a good predictor of academic achievement. The Goodenough Draw-a-Person Test of 1926 and the Goodenough-Harris revision of 1963 were intended to be culture-fair, but fairly well reflect the degree of acculturation to Western civilization (Anastasi, 1968).

Culture-fair tests are generally considered of little practical use because they are poor predictors of academic achievement. However, Evans (1974) points out a lack of logic in expecting a "culture-fair" test to predict success in a decidedly unfair society-including its educational system. He charges that psychologists defend biased instruments by pointing to their empirical predictive validity (i.e., predictability of academic achievement), and in turn educators defend their failure to teach by pointing to learners' inferior intellectual potential as assessed by psychologists' tests. Evans postulates that the lack of correlation between culture-fair tests and academic achievement could be "simply a statistical artifact

relecting the ubiquitous effects of cultural and racial bias in American society." He concludes that

one cannot logically apply clinical models based on individual deviations from a cultural norm to groups of individuals with different cultural experiences and standards. Moreover one should not even attempt to extrapolate from such models without first taking into account the situational validity of the constructs involved. (Evans, 1974)

The consensus of opinion among the majority of psychologists seems to be that even if it were possible to devise a truly culture-free test, the product would cover up so many group differences as to make predictions of success in a given society impossible (Samuda, 1975), or it would be so culturally nonspecific as to lose much information that is critically important in predicting the subject's performance (Sabatino, Kelly & Hayden, 1973). And since the attempts at developing culture-fair tests has so far resulted in tests that do show differential effects due to SES and ethnicity, this line of work had been largely abandoned until the development of civil rights legislation which demanded that instruments be nonbiased.

Culture-Specific (Pluralistic) Tests and Adjustment of Norms

Interestingly, paralleling the demise of culture-fair test development has been a theoretical push for the development of "culture-specific" or pluralistic tests which allow comparison of individuals against their own cultural reference group. Culture-specific testing is perhaps most readily seen as appropriate for those groups whose primary language is not standard English.

At a simple level, a mental ability test may be considered to be made more appropriate for a member of a minority group whose primary language is not standard English by the translation of the test into

the person's primary language. An example of this would be translation of the WISC-R into Spanish for administration to a Spanish-speaking student. At the crudest level, the test administrator would translate as the test is administered. The disadvantages to this approach are several. Comparison of the testee's performance against the test's norms becomes highly suspect because of the violation of standardized administration procedure. The quality of the translation is only as good as the extent of the administrator's bilingual skills, and the quality of the translation can vary from administration to administration. Since no testee would be subjected to the identical test from administrator to administrator, and even from administration to administration by the same examiner, development of more empirical norms to aid in interpretation is made impossible.

Test translation as an adaptive assessment strategy becomes more acceptable when a "standard" translation is made and field tested for understandability by the group for which it was developed. Using this approach all examiners use the same words when administering the test items, and replicability of administration (and consequently, transportability of the obtained results from examiner to examiner) are enhanced. There still remains the problem, however, that translation can drastically alter the difficulty level of test items, so the applicability of the standard-administration norms provided with the test remains very questionable (Sabatino et al., 1973).

Better practice for development of translated tests involves field testing of the translated test with the intended group and determining that the resulting scores reflect the distribution of

scores that theoretically should apply. Best practice might require actual renorming of the test for the population for which application is desired.

Some researchers in the field have suggested that intelligence tests should be developed relevant for each major cultural group in the United States. For example, Garcia (1981) calls for testing important mental processes for each regional, ethnic and social group using test items and methods relevant to each unique group. He suggests development of huge computer-accessed item pools with constant updating and random removal of items from the pool to generate unlimited unique tests.

Rather less ambitious is the System of Multicultural Pluralistic Assessment, or SOMPA (Mercer, 1978). Briefly, in this procedure, the student is given a standard administration of the WISC-R which is scored in the usual manner. The student's parent is then interviewed to determine the family's SES and sociocultural background. On the basis of the interview data, the child is assigned to a sociocultural group for purposes of test interpretation, and the child's obtained scores are compared to norms for his sociocultural group. Originally developed to offer an alternative to traditional mental ability testing in light of court rulings (for example, Larry P. v. Riles [1979] and Diana v. California State Board of Education [1973], stipulated settlement) holding that children from "experientially, linguistically and culturally unique" backgrounds have their potentials masked by traditional tests (Figueroa, 1979), and federal regulations for the Education of All Handicapped Children Act of 1976

(P. L. 94-142) that require that tests be "selected and administered so as not to be racially or culturally discriminatory" (34 C.F.R. §104.35; 34 C.F.R. §300.532), implementation of the SOMPA process was supposed to end the overrepresentation of minorities in special classes.

Critics of the culture-specific testing movement question the theoretical basis of such practices. Since the theoretical notion is that comparisons can only be made to persons of the same linguistic and cultural background, there would seem to be little use of attempting to generalize to out-of-culture performance, such as that in school settings (Samuda, 1975; Reschley, 1981).

The SOMPA has been available sufficiently long to have had some published research reports concerning the application of its intellectual measure, the estimated learning potential or ELP. ELPs are identical to IQs for middle and upper class white children, and relatively larger than IQs for non-majority group children, with the discrepancy growing as cultural dissimilarity increases. Kaufman (1979) questions whether the regression equations derived from the SOMPA's California norm group would generalize to children from other geographic regions, especially Southern black children. Brown (1979) criticized the technical documentation of the SOMPA and the legitimacy of its statistical assumptions. Prasse (1981) points out that the SOMPA is strongly rooted in "anti-labeling" theory, yet its only apparent application is in the labeling process. Regrettably, the SOMPA manuals and Mercer's writings provide little help in how data obtained from the test will yield more appropriate programming.

A major problem with interpretation of SOMPA profiles is determining the meaningfulness of the generated ELPs. Cleary et al. (1975) caution against the use of separate norms to award higher rankings to minority group members than would be received from mixed norms, if prediction for the total society is the goal. Like Mercer, Gerry (1973) rejects the "deficit theory" of test mean differences (i.e., the theory that something is "wrong" with the lower-scoring minority's culture to account for the lower scores), but cautions that lowering the "failing" score for minorities is much too simplistic. Fine (1975) in a book highly critical of current assessment practices offers the ultimate extreme of such a practice:

Denouncing the devastating implications of IQ tests, and the imprecision of their measurement, Professor William F. Brazziel of Virginia State College warns that a caste status has taken place in our schools An Ethnic Success Quotient is proposed by Professor Brazziel. Under this E.S.Q. test, a Richmond-born Episcopalian of English stock, from a family with an income of \$12,000, would be declared below average intelligence if his Stanford-Binet score was 120 or under. A score of 100 would place him in the low-normal or retarded class. On the other hand, a Beauford County, South Carolina, poverty child with an IQ of 100 would be placed in the bright group on the basis of the Ethnic Success Quotient concept. (p. 223)

Other recommendations to raise scores and thus avoid the danger of misidentifying retardation ("pseudoretardation") include such adjustments as defining the Binet basal as the age level immediately below the highest level at which any subtest was passed. Approaches such as this, however, are open to criticism for lack of justification for retaining the same criterion for assigning retardation ($IQ < 70$) and changing the procedure. Jastak (1949; cited in Silverstein, 1970) proposed that subjects be labeled as retarded only when scores on every task were in the retarded range. Research on this notion showed

that such a procedure made little or no difference in actual practice (Silverstein, 1970).

Piagetian Assessment Instruments

The potential usefulness of a Piagetian-based test for psychologist practitioners in schools is readily apparent. For example, the first-grader referred for problems with math could be having a problem with conservation. The student could not be expected to understand measurement if the perceived size of the thing being measured changes when moved from place to place. On the other hand, other students who "know" their math facts can work with math as rote memorization of rules without understanding what is being done. And one would not expect children functioning in a concrete operational stage to be able to handle abstract reasoning. Sharp (1972) criticizes standard intelligence tests for not being based on any elucidated theory of mental development, and cites research at the University of Montreal and the University of Geneva toward development of a Piagetian-oriented measurement instrument.

Unfortunately, Piaget's work does not readily lend itself to generation of an interpretable profile with pedagogical implications (Kohlberg, 1968). Elkind warns that much of what is represented as Piagetian is of dubious heritage:

Of particular note is the fact that Piaget nowhere points to the practical implications of his work. Indeed he seems to feel that a solid foundation of genetic research and theory is necessary before useful practical implications can be drawn. . . . This is an important point. Some educators have engaged in certain teaching practices in the name of Piaget. What should always be made clear in such cases is that it is the educator's interpretation of Piaget which is being utilized and not Piaget's own ideas about educational

practice. If Piaget has such he seldom voices them.
(Elkind in Piaget [1964/1967], pp. xv-xvi)

Changes in the Clinical Testing Interview: The Examiner and the Groundrules

The majority of recommendations designed to improve the quality of nondiscriminatory assessment of minority group students has dealt with the testing interview and the testing behavior of the examiner.

The earlier mentioned SPSSI policy statement (Fishman et al., 1964) calls for retraining of personnel to understand minority cultures in order to enhance accurate interpretation of test performance. The group also suggests some modification of test administration and procedures, but the main thrust of the statement was to interpret test results in light of influence from what are classically considered to be "extraneous" variables. The SPSSI statement implicitly accepts the cultural deficit model, and recommends continued use of Anglo tests with minorities to determine "the magnitude of the deprivation to be overcome."

Clarizio (1979) cites literature which found that using praise, extending time limits and mixing hard and easy items provide minority students with feelings of success and typically result in up to 10-15 point increases in IQ scores. Retesting minority students 1 to 2 weeks later can raise scores by up to 10 points, presumably because of a decrease in situational test anxiety. Clarizio recommends that examiners be sensitive about unexamined and unrealized ethnic stereotypes. When testing young disadvantaged students, he also recommends that the examiner allow plenty of time to make the student comfortable and that the examiner should preferably already be known and liked by

the child. This would seem to be a good way of avoiding the implicit hostility of the testing situation which is sometimes present (Flaughner, 1978).

Rewarding students for improving performance on mental ability tests has a differential effect on students from low SES families and those from middle SES families, with lower class students showing much greater improvement. Extending time limits can have a similar differential impact (Haggard, 1954).

Encouraging young students from low SES families to verbalize the steps involved in solving problems enhances their performance, according to research by Jensen (1968). Jensen hypothesizes that this results from a lessened tendency of children from poorer classes to verbalize spontaneously and thus bring their verbal ability to bear upon the task at hand.

Riessman (1974) suggests that student motivation be enhanced by creating a game situation and by giving students directed practice taking tests (Haggard, 1954). Examiners should receive careful training in achieving rapport with culturally dissimilar students. Riessman also offers the following advice:

[Test items] need not be brief; tasks showing no change with age can be utilized; time limits can be removed or lengthened; less academic, bookish problems can be employed; the thinking involved in solving the problems can be evaluated along with the accuracy of the final answer; items that discriminate between deprived and middle-class groups can be removed from the tests, just as items discriminating between the sexes are removed in test construction; emotionally loaded items can be selected and the responses to them compared with the more value-free items. (Riessman, 1974)

Regardless of what clinical exceptions are made, examiners need to be trained not to overgeneralize from the test results. A helpful

supplement to standard practice which would help avoid overgeneralization is gathering of data from the student's home environment (Gerry, 1973; Samuda, 1975; Mercer, 1978).

The literature has also focused attention to the written report which is almost always a final product of the evaluation process. The SPSSI statement urges examiners to put in cautionary statements regarding taking scores at face value when subjects' backgrounds fall outside that of the norm group. Mason (1975) goes so far as to state that psychologists should either not formally report or should at least deemphasize negative material found during assessment. He thinks that if IQ scores must be reported, the report should be limited to reporting scores as either "above average" or "not above average." Use of special education labels should be avoided. Only strengths should be referred to in reports, and, in fact, all analysis should be minimized to allow emphasis upon the recommendations section of the report.

Mason (1975) thinks that the role of the psychologist should be only to relate to teachers' behavioral findings, not clinical nomenclature or statements of underlying dynamics. Reschley (1981) calls for the psychologist's role as report writer to be deemphasized and his role as program planner/intervener to be developed. He feels that this would drastically help improve the educational consequences of being identified as handicapped, and make involvement in special education programming a worthwhile school experience.

Sabatino et al. (1973) insist that professional schools should aggressively recruit bilingual graduate students and teachers to become school psychologists.

A spokesman for those who reject the entire concept of ability testing for minority cultures, Bernal labels recommendations for such "improvements" as ethnic examiners, utilization of different testing procedures, dialectical translation, culture-specific renorming, and awarding of "bonus points" for ethnic membership as "simple minded," "condescending," and "[not questioning] the test's appropriateness" (Bernal, 1975).

Criterion Referenced Tests

Many educators, who are more concerned with the practical uses of intelligence tests rather than their theoretical implications, have called for a redefinition of the function of intelligence testing. Gordon and Terrell (1981) state that the purpose of intellectual evaluation must be changed from one of identification to pedagogy and rehabilitation. Fine (1975) says educators need to make their own criterion referenced tests and thereby avoid the dangers of reliance upon standardized tests. Brody and Brody (1976) believe that a combination of achievement tests, teacher observations and criterion referenced tests are superior to use of intelligence tests for the purpose of individualization of instruction. Cleary et al. (1975) suggest the use of cumulative diagnostic tests which involve detailed analysis of skill areas, high reliability because of large item pools, and direct connection into a self-paced instructional program.

Criterion referenced testing is attractive because it makes the tests educationally diagnostic, by revealing something of the student's educational strengths and weaknesses, and allowing the development of programs and instructional procedures matched to the particular individual's needs. Criterion referenced testing avoids any inter-individual comparisons since testee performance is compared only to an absolute criterion and the individual's distance from meeting the criterion score.

However, criterion referenced testing is not a panacea. We lack experimental evidence to show that task sequencing from criterion referenced testing actually improves the rate of skill acquisition (Samuda, 1975). And criterion referenced measures have no less potential for prejudicial assessment than do traditional psychometric instruments: "What is the criterion and who specifies the criterion to be attained?" (Drew, 1973; Samuda, 1975). Criterion referenced tests show similar levels of deficits for minorities which score low on norm referenced tests (Pyle, 1979).

Development of New Types of Tests

In 1958, Eisman urged development of a diagnostic instrument which would sample behavior in a variety of different learning conditions, since differential analysis of performance obtained in different learning situations should reveal much more about a handicapped learner's ability than a single score from an intelligence test. In 1979, Feuerstein was attempting to provide such a device.

Because the chief purchasers of tests (that is, educational institutions) are now demanding that tests provide guidance in

intervention, assessment instruments of the future need to deal with a broader concept of intelligence than scholastic predictive ability, and be more aware of building blocks to intellectual competence. Future tests need to be more readily sensitive to the effects of dissimilar cultural background (Gordon & Terrell, 1981). Other researchers feel that test scores can ultimately be related to brain functions and to specific, discrete intellectual skills (Garcia, 1981).

Use of Learning Measurements

A review of the literature reveals that surprisingly little is known about the predictive significance of individual differences in the amount of time required to reach a criterion of mastery. Most studies which have looked at the issue find a positive relationship between measured intelligence and rate of learning. However, the degree of relationship generally falls off rapidly past a certain minimal level of intellectual ability for many types of learning tasks (Eisman, 1958; Rapier, 1962).

Proponents of this line of research point out that intelligence tests measure retention and not learning. "Learning" may be defined as the "modification of behavior through experience," where experience is "the behavior elicited by definite stimulus conditions" (Simrall, 1947).

There are some theoretical and methodological difficulties associated with learning research. Theories of intelligence often provide little help as to whether MA, IQ or both should be related to learning. If MA is a measure of the developmental level of the

organism, it should delimit learning potential. On the other hand, if IQ is a life-time ratio of the rate of learning, it should be most strongly related to learning (Zeaman & House, 1967). Researchers such as Simrall (1947) conclude that "intelligence" (operationally defined by IQ) and the ability to learn are independent variables. However, Simrall and others (Woodrow, 1946; Rapier, 1962) assume that intelligence scores measure a single factor (g?) and that the IQ test is an optimal measure of that factor. To the degree that tests are not that, their conclusions will be inadequate. In general, the learning research is plagued with such methodological problems as lack of description of content of the criterion learning tasks, lack of information about the intellectual range of the subjects (but most likely of a very restricted range), and generally poor documentation of the time span allowed for subjects' learning to take place (Tilton, 1949; Rapier, 1962).

"Learning" has generally been measured in two ways (Rapier, 1962): the use of gain scores and the use of total scores. Gain scores (the final criterion score minus the beginning criterion score) are more frequently used. Gain scores are clearly a measure of improveability due to practice, but use of this technique can penalize brighter subjects by allowing insufficient ceiling on the criterion tasks and by the likelihood that the duller subjects make their gains on earlier, easier items of the criterion.

The total scores measure of learning is generally defined as the number of trials required or the number of errors made before reaching criterion. This approach has the advantage of being a more reliable

measure than the gain score approach, but has the disadvantage of being insensitive to actual learning taking place and perpetuating the entry skill level present at the beginning of practice. The entry level tends to be highly correlated with intellectual ability.

Rapier's (1962) literature review on intelligence and learning found that the frequent effect of practice is to lower the relationship between IQ and learning. One generally finds a lack of relationship between IQ and gain scores even when strong relationships exist between IQ and the initial criterion score and IQ and the final score.

When relationships between IQ and gain scores are found, learning ability measures correlate positively with IQ but at a very low level (in range of .10 - .30). The relationship reaches its maximum rather early in the school grades and there is a lack of relationship beyond the fifth grade. IQ and gain scores are more substantial for lower IQ subjects.

Zeaman and House (1967) report that IQ (with MA held constant or controlled) has no correlation with learning of classical conditioning tasks except where fine discrimination between stimuli are required. When a wide range of intellectual abilities are tested and tasks of "intermediate difficulty" are used, a low positive relationship exists between IQ (MA controlled) and performance in simple visual discrimination learning. On studies of IQ and verbal learning, Zeaman and House report that IQ and performance are clearly related where the sample's IQ range was 40 or more points, and marginal or no relationship when the sample's range was 20 to 30 points. A positive

relationship on the order of that for verbal learning was found on learning sets tasks.

Narrowed Criteria for Mental Retardation Label

In order to receive state or federal funds for provision of services to handicapped students, each local educational agency must certify that it practices non-discriminatory assessment. Tennessee State Department of Education Policy states that:

Each local school system must ensure that tests and evaluation materials meet the following criteria:

1. Are selected and administered so as not to be racially or culturally discriminatory.
2. Are provided and administered in the child's native language or other mode of communication.
3. Have been validated for the specific purpose for which they are used.
4. Are administered by trained personnel in conformance with the instructions provided by their producer.
5. Include those tailored to assess specific areas of educational need and not merely those which are designed to provide a single general intelligence quotient.
6. Are selected and administered so as to best ensure that when a test is administered to a child with impaired sensory, manual, or speaking skills, the test results accurately reflect the child's aptitude, achievement, or other level the test purports to measure. (Tennessee State Department of Education, 1983, §2.2.1)

Recognizing that "mental retardation" did not imply the comprehensive type of deficits inherent in the term "mental deficiency," the American Association of Mental Deficiency's criteria for diagnosis of mental retardation (Heber, 1961) rejected the use of adaptive behavior measures as a necessary concomitant for the diagnosis of mental retardation, since the proper adaptive behavior

domain for a school-aged youngster was seen as necessarily school adequacy. Later AAMD standards (Grossman, 1973) added impaired adaptive functioning as a part of mental retardation. Making a clean break with the historical distinction between mental retardation and mental deficiency (Sarason & Doris, 1969; Heber, 1961; Budoff, 1975), and effectively redefining mental retardation as mental deficiency, the state's Student Evaluation Manual (Tennessee State Department of Education, 1982) states that "for a child to be diagnosed as mentally retarded, impairments in intellectual functioning must co-exist with deficits in adaptive behavior." Adaptive behavior may be assessed by either systematic observations or use of commercially available standardized instruments. Either procedure may be used to assess at-home adaptive levels.

Tennessee State Board of Education (1982) Rules, Regulations, and Minimum Standards establishes as state policy that

a child who has or develops a continuing handicap in intellectual functioning and adaptive behavior which significantly impairs the ability to think and/or act and the ability to relate to and cope with the environment, is considered mentally retarded.

Functionally, diagnosis is made according to three criteria: (1) intelligence score two or more deviations "below the mean of normal distribution," (2) "significantly impaired adaptive behavior in the home or community," and (3) "significantly impaired adaptive behavior in the school, day care center, or residential program" (Tennessee State Board of Education, 1982). The psychologist is allowed to decide what constitutes "significantly impaired" behavior, but the "best practice" recommendation is to use a three percent criterion on standardized instruments (Tennessee State Department of Education,

1982). Only three years earlier no mention of any adaptive deficit was made in the state regulations for identification of mental retardation (Tennessee State Board of Education, 1979).

In the educational setting one systematic observation is required and administration of the norm-referenced measure is required where feasible. The State Board of Education (1982) requires that adaptive behavior in the home or community be assessed by use of

an appropriate individual standardized instrument to be completed with the child's principal caretaker and/or . . . systematic documented observations, impressions and development history by an appropriate specialist with the principal caretaker in the home, community, residential program or institutional setting.

Adaptive behavior in the educational setting is to be assessed by

systematic documented observations by an appropriate specialist which compare the child with other children of his/her age group, and when feasible . . . scores on an appropriate individual standardized instrument to be completed with the principal teacher of the child or the child.

The Student Evaluation Manual recommends use of three instruments for assessment of adaptive behavior: the Adaptive Behavior Inventory for Children, which is part of the SOMPA (Mercer, 1978); the AAMD Adaptive Behavior Scale School Edition (Lambert, Windmiller, Cole & Figueroa, 1975); and the Vineland Social Maturity Scale with normative scoring as adapted by Silverstein (1971).

Whether the use of adaptive measures and observational data will result in less discriminatory assessment of non-majority group members is debatable. Lambert (1978) found that less than ten percent of observable classroom behaviors had to do with activities associated with intelligence. Adaptive behavior scales are extremely culturally loaded (Clarizio, 1979).

II. METHOD

Subjects and Setting

Students enrolled in two rural elementary schools in East Tennessee participated in this study. Each school served students from kindergarten through the eighth grade. One school (School A) had an enrollment of approximately 600 students; School B had an enrollment of approximately 680. The student feeder population for each school came from farms or a small rural town with population less than 1000.

Forty students were selected to fit a two by two design. Ten subjects were obtained for each of the following cells: 6-year-old males, 6-year-old females, 11-year-old males, and 11-year-old females. All subjects were white, native-born Americans.

To assure that academic instruction was being offered in material to which the achievement tests utilized in this study might be expected to be sensitive, a decision was made not to select six-year-olds from kindergarten classes. Tables 2 and 3 present summaries of the grade placements and ages of the subjects.

Subjects were selected from a roster of students who received free school lunches based on their families' reported incomes. Since the information obtained on the families' applications was known to be subject to verification, this assured selection of students from families meeting federally-established criteria for low income status.

The parents of all students which attended School A and which met the selection criteria were invited to allow their children to participate. Students from School B were selected and invited to participate in sufficient numbers to complete the necessary sample sizes. All

Table 2. Grade Placement for subjects in sample

	Grade			
	1	5	6	7
<u>Age 6</u>				
Male	10			
Female	10			
<u>Age 11</u>				
Male		2	7	1
Female		2	7	1

Table 3. Summary of age data for subjects

	Youngest ^a	Oldest ^a	Median ^a	Mean ^b	Standard Deviation ^b
<u>Age 6</u>					
Male	6-2	6-11	6-7	6.55	0.24
Female	6-0	6-11	6-4.5	6.43	0.33
<u>Age 11</u>					
Male	11-0	11-11	11-6.5	11.53	0.36
Female	11-1	11-11	11-4.5	11.46	0.33

^aReported as months - years, so that 11-6 equals 11 years 6 months.

^bReported as decimal value in years, so that 11.50 equals 11 years 6 months. One month equals 0.083 year.

eligible six-year-old females were invited from this school. Invitations to subjects in the other three cells were made by random selection. Table 4 presents a summary of information relating to which students were selected from each school.

Instruments

Four standardized instruments were administered to each subject. These instruments were the Wechsler Intelligence Scale for Children--Revised (Wechsler, 1974), the Kaufman Assessment Battery for Children (Kaufman and Kaufman, 1983), the Peabody Individual Achievement Test (Dunn and Markwardt, 1970), and the Wide Range Achievement Test (Jastak & Jastak, 1978). The mathematics, reading recognition and reading comprehension subtests were administered from the PIAT. The reading and arithmetic subtests were administered from the WRAT.

Wechsler Intelligence Scale for Children--Revised. The WISC-R is a well-established measure of intellectual ability which is extensively described in the psychological literature. The WISC-R provides three global standard scores including the full scale quotient (FS-IQ), verbal quotient (V-IQ) and performance quotient (P-IQ). Each scale is normed with a mean of 100 and standard deviation of 15. The WISC-R also provides 12 subtests although only 10 are used in determination of the three global measures. The test is normed for children between the ages of 6 and 16, and requires approximately 45 minutes to administer.

System of Multicultural Pluralistic Assessment. The SOMPA is offered as a comprehensive system of measuring a student's cognitive abilities, sensorimotor abilities and level of adaptive behavior skills.

Table 4. Degree to which sample saturated population of interest at each school

	School A			School B		
	Eligible	Invited	Included	Eligible	Invited	Included
<u>Age 6</u>						
Male	12	12	7	8	5	3
Female	4	4	4	7	7	6
<u>Age 11</u>						
Male	9	9	8	5	3	2
Female	12	12	8	4	2	2

Sensorimotor assessment is accomplished by administration of the Bender Visual Motor Gestalt Test (Bender, 1975), measuring a student's performance on several physical dexterity tasks, and by recording information about weight, height, visual acuity and auditory acuity. Another component is the Adaptive Behavior Inventory for Children, a series of graded questions involving the student's interaction with the culture outside the school. A Health Inventory Questionnaire gathers information from the child's mother about the mother's pregnancy and the child's history of illnesses and accidents. A parent interview also provides information for fitting the child's home environment into the SOMPA's schedule of demographic scales.

The SOMPA's demographic classes are defined according to the variables of family size, family structure, socioeconomic status, and urban acculturation. Family structure refers to the degree to which the nuclear family is intact. Urban acculturation involves the family's sense of locus of control, social involvement in the community, education and rural/urban rearing of the parents, and language spoken in the home.

While obtaining SOMPA intellectual ability scores required no additional effort from the subjects beyond taking the WISC-R (the other evaluation components were not administered), each subject's parent was interviewed by the examiner to complete the Sociocultural Scale. This completed scale was later used in the data analysis stage to determine the statistical parameters necessary to derive the SOMPA's statistic of estimated learning potential (the ELP) from the WISC-R IQ. The ELP is approximately equal to the IQ plus the difference between the expected

IQ of the subject's demographic class and 100. If the expected class IQ is greater than 100, the ELP equals the obtained IQ. Separate norms are used for white, black and Hispanic subjects. The SOMPA is normed for children aged 5 to 11; younger students substitute their Wechsler Preschool and Primary Scale of Intelligence (Wechsler, 1967) scores for those of the WISC-R.

Kaufman Assessment Battery for Children. The K-ABC yields four global standard scores plus a number of subtest scores. The global scores are Simultaneous Processing, Sequential Processing, the global intellectual measure of the Mental Processing Composite (MPC) plus an overall Achievement score. Means and deviations are identical to those of the WISC-R. Students are given between 7 to 13 of the K-ABC's 16 subtests according to students' age. The test is normed for children aged $2\frac{1}{2}$ to $12\frac{1}{2}$ years. Although not utilized in this study, supplemental sociocultural norms are provided for blacks and whites according to parents' education (less than, equal to, or greater than 12 years). Administration takes about 75 minutes for students in the age span of this experiment.

The K-ABC is built upon an information-processing approach to intellectual functioning. In contrast to the content-approach of the WISC-R (and thus, by default, the SOMPA), the emphasis of the K-ABC is upon whether the stimuli are managed one at a time or simultaneously, regardless of content. Sequential scale items require subjects to solve problems by dealing with stimuli in sequential or serial order. Simultaneous items involve simultaneous manipulation of information, obtaining a gestalt, and synthesizing diverse input. The Mental

Processing Composite is a global index which pools the Sequential and Simultaneous scores. The global Achievement index is based on students' performance on traditional academic tasks.

Peabody Individual Achievement Test. The PIAT is a well-established, individually administered academic achievement test normed for students aged 5½ to 18½. Subtests measure reading decoding, reading comprehension, arithmetic, spelling and general knowledge. Only the first three subtests were administered to subjects in this study. While standard scores normed on grade placement are available, only standard scores normed on age were considered in this study.

Wide Range Achievement Test. The WRAT is an individually administered test of reading decoding, arithmetic computation and spelling, the latter not administered to subjects in this study. With norms for individuals aged 5½ through adulthood, the WRAT is an immensely popular test when rapid results from individual administrations are required.

The author-test administrator had university training and several years experience in administration and interpretation of the achievement tests and the WISC-R. He was trained in May, 1980, by the SOMPA's primary author, Dr. Jane R. Mercer, and is a certified SOMPA examiner and interpreter. This training took place in a University of California, Riverside, extension course. The author had also independently studied and administered about 15 K-ABC evaluations prior to beginning the present study.

Procedure

Early in the school year, each subject was administered the four standardized instruments. The WISC-R and K-ABC were always administered before the achievement tests. The two intelligence tests and the two achievement tests were otherwise administered in a latin-square counter-balanced order to rule out practice effects.

All tests were administered by one examiner (the author). To control for extraneous order effects, the order of the cells from which the subjects were tested was counterbalanced. The first round of test administration began on the nineteenth day of school and continued over the next 31 school days. Most subjects took all four tests in one day (approximately two to three hours). If the examiner detected that a subject's performance was beginning to wane, testing was scheduled to be completed at a later time.

Between the first and second rounds of subject testing, each subject's parent was administered the Sociocultural scale of the SOMPA. Because three families had two participants involved in the study, 37 separate parent interviews were conducted.

Beginning in mid-April and over the next 6 weeks, each subject was retested. This second round of testing consisted of readministration of only the two achievement tests given earlier. The 6-year-old sample was also administered the Achievement subtests only of the K-ABC. Subjects were retested in the order in which they were originally seen, to allow for approximately equal periods of time between the first and second testings. The mean period of time between tests was 125.75 days, with a range of 114 to 131 days and the standard deviation equal to 2.84.

To minimize the possible effects of examiner bias during the data gathering phase, no protocols were scored until all data were collected. A second measure was made to check the reliability of protocol scoring. A random sample, consisting of twenty percent of the test protocols, were photocopied and scored by another certified professional qualified to administer these tests.

Data Analysis

Standard scores were used in all analyses except where noted otherwise. When two or more variables were contrasted, a multiple-factor analysis of variance (ANOVA) with repeated measures on one factor (Tests) was employed.

Ability tests were evaluated for their accuracy in prediction of achievement test scores based on the ratio of predicted achievement/actual achievement. The ability test standard score was substituted for the predicted achievement score. This direct substitution procedure is consistent with the regulations of many state departments of education (including Tennessee's) in the diagnosis of specific learning disabilities (Naglieri, in press a; Tennessee State Department of Education, 1982), in which a student's academic achievement is compared directly to his or her full scale intelligence score to derive a discrepancy index. The assumption of this regulatory model is that the relationship between achievement standard scores and any acceptable intelligence test's standard score is

$$\text{Achievement}_{\text{Expected}} = 1 \times \text{IQ}_{\text{Full Scale}}$$

In its implementation, a perfect match between ability quotient and achievement quotient is assigned a unity value. "Overpredictions," where obtained achievement is less than that predicted, have values greater than unity and increasing with the percentage of difference. "Underpredictions" obtain corresponding values.

To minimize the loss of power because of methodological considerations, only the statistics of interest because of planned comparisons are evaluated for their significance. Table 5 presents the data to which these analyses apply. There were fourteen variables of interest plus the overall ANOVA of unaltered global scores of the three intelligence tests, and hence, fifteen ANOVA analyses to be determined. Solving the equation

$$\text{Probability (Type I Error)} = 1 - (1 - \alpha)^T$$

where $T = 15$ and Probability (Type I Error) = .05 or .01, α values of .0034 and .001, respectively, were obtained. The value of F cutting off the upper .0034 or .001 portion of a distribution with the proper degrees of freedom was then used in the tests for significance.

Hypotheses

For purposes of statistical analysis, the following hypotheses were utilized:

Hypothesis 1: The WISC-R, SOMPA and K-ABC mean scores will not differ significantly.

Hypothesis 2: The WISC-R, SOMPA and K-ABC generate scores that are the same for students in this low income, white sample.

Table 5. Achievement data collected according to sample^a

	Achievement Test ^b				
	Pretest K-ABC	Pretest PIAT	Pretest WRAT	Posttest PIAT	Posttest WRAT
<u>Total Sample</u> (N = 40)	Math Decoding	Math Decoding		Math Decoding Comp.	
<u>6-year-old Sample</u> (N = 20)			Math Decoding		Math Decoding
<u>11-year-old Sample</u> (N = 20)	Comp.		Math (GE)		Math (GE)

^aData are standard scores except as noted (GE = grade equivalent scores).

^bSubtests are classified according to the skill measured rather than formal name. The K-ABC Decoding subtest, PIAT Reading Recognition subtest, and WRAT Reading subtest are accordingly all designated as "Decoding." Reading comprehension subtest scores are designated as "Comp."

Hypothesis 3: The WISC-R, SOMPA and K-ABC full scale standard scores will not differ in their accuracy in simple ratio prediction of achievement.

Hypothesis 4: There will be no interaction between the variable of subject age and the intellectual tests' accuracy in simple ratio prediction of achievement.

Hypothesis 5: There will be no interaction between the variable of subject sex and the intellectual tests' accuracy in simple ratio prediction of achievement.

III. RESULTS

Reliability

Reliability of scoring the test protocols was established using the method recommended by Winer (1971, p. 283). This method employs a single-factor ANOVA with repeated measures on one factor to estimate the reliability of measurements. The following estimates of the reliability of a single scorer were obtained: WISC-R, .984; K-ABC, .993; PIAT, .992; and WRAT, .996. No separate estimate of reliability was made for the SOMPA, as SOMPA scores are obtained by statistical manipulation of WISC-R global scores. However, no errors in computation were found when double-checking the calculations. In summary, the above estimates of scorer reliability were very adequate for the purposes of this research.

Means, standard deviations and ranges of standard scores obtained by the sample on the WISC-R, SOMPA, K-ABC and PIAT are presented in Table 6. Restricted range of scores is frequently a problem in research of this type, but the obtained ability scores from this sample spread over about 4 standard deviations. The PIAT achievement scores were spread over about 3 deviations. The full scale ability test means were about .4 - .6 deviations below the normative mean, which is consistent with what would be expected of a rural and low SES population.

Comparison of mean ability scores (Hypothesis 1)

Examination of the data was made to determine if the three intelligence tests yielded different sample means for these subjects. Rejection of Hypothesis 1 was the expected finding, since subjects

Table 6. Means, standard deviations and ranges of WISC-R, SOMPA, K-ABC and PIAT scores (N = 40)

	Descriptive Statistics			
	Mean	Standard Deviation	Minimum Score	Maximum Score
<u>WISC-R</u>				
Verbal Scale	90.90	13.24	65	128
Performance Scale	92.98	12.54	69	120
Full Scale	91.10	12.57	64	121
<u>SOMPA</u>				
Verbal Scale	96.85	14.28	73	136
Performance Scale	94.45	12.06	69	120
Full Scale	93.93	13.08	66	126
<u>K-ABC</u>				
Simultaneous Processing Scale	91.33	12.70	65	120
Sequential Processing Scale	95.38	15.91	56	126
Mental Processing Scale	92.15	14.14	60	123
Achievement Scale	91.25	11.78	66	125
<u>PIAT (Pretest)</u>				
Reading Recognition	97.28	9.83	72	119
Reading Comprehension	96.83	9.52	72	118
Arithmetic	95.30	10.67	74	120
<u>PIAT (Posttest)</u>				
Reading Recognition	101.28	11.07	80	123
Reading Comprehension	100.03	11.00	75	128
Arithmetic	96.18	11.28	79	120

were selected to represent a lower SES status, and the SOMPA and K-ABC were designed to minimize the statistical relationship between SES and school ability measures. A three-factor ANOVA with repeated measures on one factor (Tests) was conducted, as discussed in Winer (1971). Table 7 reports the ANOVA summary for this analysis. Neither primary nor interaction effects were found, and the hypothesis was not rejected.

Comparison of distributions of scores (Hypothesis 2)

Correlations among the WISC-R, SOMPA and K-ABC global scores are presented in Table 8. To test Hypothesis 2, three correlations were of interest: Full Scale WISC-R vs. Full Scale SOMPA (.96), Full Scale WISC-R vs. the MPC of the K=ABC (.69), and Full Scale SOMPA vs. the MPC of the K-ABC (.65). Obtained t values against the alternative that the true correlations were zero were, respectively, 21.14, 5.88 and 5.27. Each of these correlations were determined to be significant, with $p < .01$ after α was adjusted for repeated testing. Critical $t_{.01} = 3.58$, corresponding to .001 of a distribution with 38 degrees of freedom. Thus Hypothesis 2 was not rejected.

Accuracy in prediction (Hypotheses 3, 4 and 5)

Correlation coefficients among the ability tests and the K-ABC achievement subtests plus the PIAT are presented in Table 9 (total sample), Table 10 (6-year-old sample) and Table 11 (12-year-old sample). Table 9 does not report correlation coefficients involving the WRAT. This is because an error in data collection resulted in an inability to generate standard scores from the WRAT for the

Table 7. Summary of analysis of variance for sex x age x test
(WISC-R Full Scale IQ, SOMPA Full Scale ELP and
K-ABC MPC)

Source	SS	df	MS	F
<u>Between Subjects</u>	17,720.59	39		
Sex	190.01	1	190.01	0.39
Age	0.41	1	0.41	0.00
Sex x Age	2.41	1	2.41	0.00
Subj. w. groups (error [between])	17,527.77	36	486.88	
<u>Within Subjects</u>	3,606.00	80		
Tests	163.12	2	81.56	1.76
Sex x Tests	63.12	2	31.56	0.68
Age x Tests	37.62	2	18.81	0.41
Sex x Age x Tests	2.62	2	1.31	0.03
Tests x Subj. w. groups (error [within])	3,339.53	72	46.38	

Note: No significant ratios were found with $p < .05$.

Table 8. Pearson product-moment correlations among the WISC-R, SOMPA and K-ABC global scores (N = 40)^a

	P-IQ	FS-IQ	V-ELP	P-ELP	FS-ELP	SEQ	SIM	MPC	Ach't
<u>WISC-R</u>									
Verbal IQ	.66	.93	.95	.60	.92	.43	.52	.53	.83
Performance IQ (P-IQ)	--	.89	.54	.94	.81	.53	.80	.74	.64
Full Scale IQ (FS-IQ)		--	.84	.82	.96*	.53	.71	.69*	.82
<u>SOMPA</u>									
Verbal ELP (V-ELP)			--	.51	.91	.42	.45	.48	.77
Performance ELP (P-ELP)				--	.80	.45	.73	.66	.54
Full Scale ELP (FS-ELP)					--	.53	.66	.65*	.79
<u>K-ABC</u>									
Sequential Scale (SEQ)						--	.66	.89	.65
Simultaneous Scale (SIM)							--	.93	.59
Mental Processing Scale (MPC)								--	.67
Achievement (Ach't)									--

^aOnly the three marked correlations were tested for significance (see text).

*p < .01 as adjusted for repeated tests.

Table 9. Correlations among WISC-R, SOMPA, K-ABC and PIAT (N = 40)

	WISC-R FS-IQ	SOMPA FS-ELP	K-ABC MPC
<u>K-ABC (Pretest)</u>			
Reading Decoding ^a	.64	.59	.51
Mathematics	.84	.84	.75
Global Achievement	.82	.79	.67
<u>PIAT (Pretest)</u>			
Reading Recognition	.71	.65	.59
Reading Comprehension	.70	.64	.57
Mathematics	.37	.34	.41
<u>PIAT (Posttest)</u>			
Reading Recognition	.70	.63	.54
Reading Comprehension	.57	.51	.47
Mathematics	.69	.67	.60

^aThe K-ABC Reading Comprehension subtest is not administered to six-year-olds. Correlations for the 11-year-old sample are found in Table 11.

Table 10. Correlations among WISC-R, SOMPA, and K-ABC full scale scores vs. achievement test scores for 6-year-old sample (N = 20 except as noted)

	WISC-R	SOMPA	K-ABC
<u>K-ABC (Pretest)</u>			
Reading Decoding	.77	.75	.63
Mathematics	.75	.76	.76
Global Achievement	.78	.78	.64
<u>PIAT (Pretest)</u>			
Reading Recognition	.76	.73	.67
Reading Comprehension	.74	.69	.65
Mathematics	.78	.71	.70
<u>WRAT (Pretest)</u>			
Reading	.74	.75	.41
Arithmetic	.87	.87	.71
<u>K-ABC (Posttest)</u>			
Reading Decoding	.68	.67	.48
Reading Comp. (N = 17)	.65	.60	.31
Mathematics	.79	.83	.77
Global Achievement	.81	.76	.63
<u>PIAT (Posttest)</u>			
Reading Decoding	.68	.61	.61
Reading Comprehension	.74	.65	.56
Mathematics	.56	.56	.76
<u>WRAT (Posttest)</u>			
Reading	.71	.66	.59
Arithmetic	.73	.77	.70

Table 11. Correlations among WISC-R, SOMPA, and K-ABC full scale scores vs. achievement test scores for 11-year-old sample (N = 20 and data are standard scores except as noted; GE = grade equivalent)

	WISC-R	SOMPA	K-ABC
<u>K-ABC (Pretest)</u>			
Reading Decoding	.64	.55	.44
Reading Comprehension	.81	.80	.65
Mathematics	.96	.94	.75
Global Achievement	.89	.84	.71
<u>PIAT (Pretest)</u>			
Reading Recognition	.69	.61	.55
Reading Comprehension	.66	.60	.51
Mathematics	.85	.79	.82
<u>WRAT (Pretest)</u>			
Reading (GE)	.70	.66	.55
Arithmetic (GE)	.87	.83	.74
<u>PIAT (Posttest)</u>			
Reading Decoding	.76	.69	.55
Reading Comprehension	.45	.40	.50
Mathematics	.80	.75	.49
<u>WRAT (Posttest)</u>			
Reading (GE) (N = 15)	.68	.59	.44
Arithmetic (GE) (N = 19)	.69	.60	.42

11-year-old sample. However, Table 10 does present the WRAT (standard scores) correlations for the 6-year-old sample and Table 11 presents the WRAT (grade equivalent) correlations for the 11-year-old sample.

No correlation is reported in Table 9 involving the K-ABC Reading Comprehension subtest. This subtest is not normed for students under seven years old, so the subtest was administered only to the 11-year-old half of the total sample. The relevant correlations for the older sample are presented in Table 11, however.

The correlations in Tables 9, 10 and 11 are presented for interest and description only. No tests of significance were made because of the methodological considerations already described.

Concurrent achievement data. Tables 12 - 16 report ANOVA summaries involving reading achievement "pretests" (i.e., concurrent administrations) as described below:

Table 12: Ability tests vs. K-ABC Decoding

Table 13: Ability tests vs. PIAT Reading Recognition

Table 14: Ability tests vs. WRAT Reading (6-year-old data only)

Table 15: Ability tests vs. WRAT Reading (GE scores; 11-year-old data only)

Table 16: Ability tests vs. K-ABC Reading Comprehension
(11-year-old data only)

Tables 17 - 20 report ANOVA summaries involving mathematics achievement pretests as described below:

Table 17: Ability tests vs. K-ABC Arithmetic

Table 18: Ability tests vs. PIAT Arithmetic

Table 12. Summary of analysis of variance for sex x age x ability tests' concurrent prediction of K-ABC Reading Decoding subtest scores

Source	SS	df	MS	F
<u>Between Subjects</u>	21,949.63	39		
Sex	2,255.43	1	2,255.43	5.51
Age	4,317.76	1	4,317.76	10.56
Sex x Age	653.64	1	653.64	1.60
Subj. w. groups (error [between])	14,722.81	36	408.97	
<u>Within Subjects</u>	4,385.40	80		
Tests	204.78	2	102.39	1.82
Sex x Tests	77.35	2	38.67	0.69
Age x Tests	44.90	2	22.45	0.40
Sex x Age x Tests	9.41	2	4.70	0.08
Tests x Subj. w. groups (error [within])	4,048.96	72	56.24	

Note: No significant ratios were found with $p < .05$.

Table 13. Summary of analysis of variance for sex x age x ability tests' concurrent prediction of PIAT Reading Recognition subtest scores

Source	SS	df	MS	F
<u>Between Subjects</u>	9,890.32	39		
Sex	1,242.87	1	1,242.87	5.26
Age	5.88	1	5.88	0.02
Sex x Age	132.37	1	132.37	0.56
Subj. w. groups (error [between])	8,509.20	36	236.37	
<u>Within Subjects</u>	3,900.27	80		
Tests	183.13	2	91.56	1.83
Sex x Tests	65.93	2	32.96	0.66
Age x Tests	40.91	2	20.45	0.41
Sex x Age x Tests	5.75	2	2.88	0.06
Tests x Subj. w. groups (error [within])	3,604.55	72	50.06	

Note: No significant ratios were found with $p < .05$.

Table 14. Summary of analysis of variance for sex x ability tests' concurrent prediction of WRAT Reading subtest scores (6-year-old data only)

Source	SS	df	MS	F
<u>Between Subjects</u>	5,204.57	19		
Sex	107.11	1	107.18	0.38
Subj. w. groups (error [between])	5,097.46	18	283.19	
<u>Within Subjects</u>	2,303.04	40		
Tests	103.51	2	51.75	0.86
Sex x Tests	26.75	2	13.37	0.22
Tests x Subj. w. groups (error [within])	2,172.79	36	60.36	

Note: No significant ratios were found with $p < .05$.

Table 15. Summary of analysis of variance for sex x ability tests' concurrent prediction of WRAT Reading subtest scores (GE scores; 11-year-old data only)

Source	SS	df	MS	F
<u>Between Subjects</u>	18,735.44	19		
Sex	4,315.45	1	4,315.45	5.39
Subj. w. groups (error [between])	14,419.99	18	801.11	
<u>Within Subjects</u>	1,613.67	40		
Tests	111.60	2	55.80	1.37
Sex x Tests	39.10	2	19.55	0.48
Tests x Subj. w. groups (error [within])	1,462.97	36	40.64	

Note: No significant ratios were found with $p < .05$.

Table 16. Summary of analysis of variance for sex x ability tests' concurrent prediction of K-ABC Reading Comprehension subtest scores (11-year-old data only)

Source	SS	df	MS	F
<u>Between Subjects</u>	4,046.32	19		
Sex	416.64	1	416.64	2.07
Subj. w. groups (error [between])	3,629.68	18	201.65	
<u>Within Subjects</u>	1,709.76	40		
Tests	113.81	2	56.90	1.32
Sex x Tests	45.06	2	22.53	0.52
Tests x Subj. w. groups (error [within])	1,550.89	36	43.08	

Note: No significant ratios were found with $p < .05$.

Table 17. Summary of analysis of variance for sex x age x ability tests' concurrent prediction of K-ABC Arithmetic subtest scores

Source	SS	df	MS	F
<u>Between Subjects</u>	6,211.43	39		
Sex	38.84	1	38.84	0.25
Age	162.33	1	162.33	1.05
Sex x Age	420.15	1	420.15	2.71
Subj. w. groups (error [between])	5,590.12	36	155.28	
<u>Within Subjects</u>	4,151.07	80		
Tests	187.71	2	93.85	1.77
Sex x Tests	83.62	2	41.81	0.79
Age x Tests	57.95	2	28.97	0.55
Sex x Age x Tests	10.32	2	5.16	0.10
Tests x Subj. w. groups (error [within])	3,811.46	72	52.94	

Note: No significant ratios were found with $p < .05$.

Table 18. Summary of analysis of variance for sex x age x ability tests' concurrent prediction of PIAT Arithmetic subtest scores

Source	SS	df	MS	F
<u>Between Subjects</u>	6,070.66	39		
Sex	158.54	1	158.51	0.99
Age	143.42	1	143.42	0.90
Sex x Age	0.30	1	0.30	0.00
Subj. w. groups (error [between])	5,768.40	36	160.23	
<u>Within Subjects</u>	4,187.61	80		
Tests	194.38	2	97.19	1.81
Sex x Tests	82.10	2	41.05	0.76
Age x Tests	38.44	2	19.22	0.36
Sex x Age x Tests	4.09	2	2.05	0.04
Tests x Subj. w. groups (error [within])	3,868.60	72	53.73	

Note: No significant ratios were found with $p < .05$.

Table 19. Summary of analysis of variance for sex x ability tests' concurrent prediction of WRAT Arithmetic subtest scores (6-year-old data only)

Source	SS	df	MS	F
<u>Between Subjects</u>	1,671.42	19		
Sex	12.52	1	12.52	0.14
Subj. w. groups (error [between])	1,658.90	18	92.16	
<u>Within Subjects</u>	2,185.62	40		
Tests	105.50	2	52.75	1.77
Sex x Tests	36.57	2	18.28	0.32
Tests x Subj. w. groups (error [within])	2,043.55	36	56.77	

Note: No significant ratios were found with $p < .05$.

Table 20. Summary of analysis of variance for sex x ability tests' concurrent prediction of WRAT Arithmetic (GE) subtest scores (11-year-old data only)

Source	SS	df	MS	F
<u>Between Subjects</u>	10,848.04	19		
Sex	2,054.18	1	2,054.18	4.20
Subj. w. groups (error [between])	8,793.86	18	488.55	
<u>Within Subjects</u>	3,152.18	40		
Tests	202.66	2	101.33	1.27
Sex x Tests	87.67	2	43.84	0.55
Tests x Subj. w. groups (error [within])	2,861.85	36	79.50	

Note: No significant ratios were found with $p < .05$.

Table 19: Ability tests vs. WRAT Arithmetic (6-year-old data only)

Table 20: Ability tests vs. WRAT Arithmetic (GE scores; 11-year-old data only)

As may be seen by examination of these tables, analyses involving the total sample (both age groups) were accomplished by a 3-way ANOVA (Sex x Age x Tests, with repeated measures on the third factor). In the cases where data was not complete for both age groups (analyses involving K-ABC reading comprehension scores or WRAT scores), a 2-way ANOVA (Sex x Tests, with repeated measures on the second factor) was used. In either case, critical F ratios reflect correction for multiple testing. These values were $F_{.05}(1,36) = 10.65$, $F_{.05}(2,72) = 6.52$, $F_{.05}(1,18) = 12.55$, $F_{.05}(2,36) = 6.87$.

These ANOVA summaries are notable for the dearth of significant findings arising from the statistical analyses. The only near-significant F ratio is found in Table 12, which reflected an interesting relationship between age and accuracy in prediction of students' skills in reading word lists on the Kaufman Decoding subtest. Inspection of the data indicated that the younger sample's achievement scores were overpredicted by the ability tests. The 6-year-old sample did not read as well as the ability tests had predicted, while the 11-year-old sample tended to read better than was predicted. "As well as" and "better than" refer to comparisons of standard scores, and the F ratio was not actually significant.

In no case did any of these ANOVA procedures reveal differential success among the three ability tests in making concurrent predictions of achievement.

Predictive achievement data. Tables 21 - 23 report ANOVA summaries involving reading achievement "posttests" (i.e., administrations seven months after administration of the ability tests) as described below.

Table 21: Ability tests vs. PIAT Reading Recognition

Table 22: Ability tests vs. PIAT Reading Comprehension

Table 23: Ability tests vs. WRAT Reading (6-year-old data only)

Tables 24 - 26 report ANOVA summaries involving mathematics achievement pretests as described below:

Table 24: Ability tests vs. PIAT Arithmetic

Table 25: Ability tests vs. WRAT Arithmetic (6-year-old data only)

Table 26: Ability tests vs. WRAT Arithmetic (GE scores; 11-year-old data only)

Similar to the preceeding concurrent achievement analyses, not one of these six ANOVAs yields a significant F ratio. Not one of the F ratios for the ability test factor even approach significance.

In summary, the data from this sample do not allow rejection of Hypothesis 3, 4 or 5.

Table 21. Summary of analysis of variance for sex x age x ability tests' future prediction of PIAT Reading Recognition subtest scores

Source	SS	df	MS	F
<u>Between Subjects</u>	10,428.15	39		
Sex	757.23	1	757.23	3.28
Age	962.52	1	962.52	4.17
Sex x Age	405.61	1	405.61	1.76
Subj. w. groups (error [between])	8,302.80	36	230.63	
<u>Within Subjects</u>	3,557.27	80		
Tests	169.41	2	84.71	1.86
Sex x Tests	58.31	2	21.32	0.64
Age x Tests	42.64	2	21.32	0.47
Sex x Age x Tests	4.45	2	2.22	0.05
Tests x Subj. w. groups (error [within])	3,282.46	72	45.59	

Note: No significant ratios were found with $p < .05$.

Table 22. Summary of analysis of variance for sex x ability tests' future prediction of PIAT Reading Comprehension subtest scores

Source	SS	df	MS	F
<u>Between Subjects</u>	14,021.25	39		
Sex	297.16	1	297.16	0.92
Age	1,924.12	1	1,924.12	5.98
Sex x Age	215.62	1	215.62	0.67
Subj. w. groups (error [between])	11,584.35	36	321.79	
<u>Within Subjects</u>	3,555.70	80		
Tests	174.58	2	87.29	1.91
Sex x Tests	55.29	2	27.65	0.60
Age x Tests	29.52	2	14.76	0.32
Sex x Age x Tests	4.74	2	2.37	0.05
Tests x Subj. w. groups (error [within])	3,291.57	72	45.72	

Note: No significant ratios were found with $p < .05$.

Table 23. Summary of analysis of variance for sex x ability tests' future prediction of WRAT Reading subtest scores (6-year-old data only)

Source	SS	df	MS	F
<u>Between Subjects</u>	5,956.96	19		
Sex	255.49	1	255.49	0.81
Subj. w. groups (error [between])	5,701.47	18	316.75	
<u>Within Subjects</u>	1,873.59	40		
Tests	92.42	2	46.21	0.94
Sex x Tests	15.91	2	7.96	0.16
Tests x Subj. w. groups (error [within])	1,765.27	36	49.04	

Note: No significant ratios were found with $p < .05$.

Table 24. Summary of analysis of variance for sex x age x ability tests' future prediction of PIAT Arithmetic subtest scores

Source	SS	df	MS	F
<u>Between Subjects</u>	12,337.13	39		
Sex	145.85	1	145.85	0.44
Age	170.50	1	170.50	0.51
Sex x Age	58.54	1	58.54	0.18
Subj. w. groups (error [between])	11,962.24	36	332.28	
<u>Within Subjects</u>	4,174.06	80		
Tests	183.30	2	91.65	1.73
Sex x Tests	105.07	2	52.53	0.99
Age x Tests	73.05	2	36.53	0.69
Sex x Age x Tests	4.06	2	2.03	0.04
Tests x Subj. w. groups (error [within])	3,808.58	72	52.90	

Note: No significant ratios were found with $p < .05$.

Table 25. Summary of analysis of variance for sex x ability tests' future prediction of WRAT Arithmetic subtest scores (6-year-old data only)

Source	SS	df	MS	F
<u>Between Subjects</u>	2,788.73	19		
Sex	21.34	1	21.34	0.14
Subj. w. groups (error [between])	2,767.39	18	153.74	
<u>Within Subjects</u>	1,962.02	40		
Tests	95.46	2	47.73	0.93
Sex x Tests	27.75	2	13.88	0.27
Tests x Subj. w. groups (error [within])	1,838.81	36	51.08	

Note: No significant ratios were found with $p < .05$.

Table 26. Summary of analysis of variance for sex x ability tests' future prediction of WRAT Arithmetic (GE) subtest scores (11-year-old data only)

Source	SS	df	MS	F
<u>Between Subjects</u>	32,950.02	19		
Sex	356.94	1	356.94	0.20
Subj. w. groups (error [between])	32,593.08	18	1,810.73	
<u>Within Subjects</u>	2,633.55	40		
Tests	207.84	2	103.92	1.61
Sex x Tests	96.85	2	48.42	0.75
Tests x Subj. w. groups (error [within])	2,328.86	36	64.69	

Note: No significant ratios were found with $p < .05$.

IV. DISCUSSION

The similarity of the WISC-R FS-IQ, SOMPA FS-IQ, and K-ABC MPC mean scores found in this study suggests that practicing psychologists may expect the three instruments to yield similar scores for students matching the demographic characteristics of this sample. Also, neither subjects' age or sex had a significant impact on their functioning of the tests.

Development of the SOMPA arose from a position that the degree to which the WISC-R and Stanford-Binet were culturally laden interfered with their utility as measurements of academic potential. By adjusting obtained WISC-R V-IQ, P-IQ and FS-IQ scores based on socio-cultural characteristics, an attempt is made to equalize for differences in educational opportunity, motivation and history of rewards for academic achievement. Much has been written about the theoretical basis of the SOMPA, both pro (Mercer, 1978, 1979a, 1979b, 1981; Mercer and Lewis, 1981) and con (Kaufman, 1979; Brown, 1979; Prasse, 1981), yet few studies are in the literature regarding empirical characteristics of the test. Oakland (1983) reports that WISC-R FS-IQ scores were raised about 5 points by SOMPA conversions for lower class white subjects. More importantly, however, Oakland found that ELPs improved neither concurrent nor future (3 years later) predictions of academic success. WISC-R FS-IQs were superior to the SOMPA FS-ELPs for this purpose.

In the present study, using SOMPA procedures made only tiny differences in intellectual ability scores. For this sample, then, which was low income, white, rural and with many "middle class"

characteristics (English-speaking, high family cohesiveness), the SOMPA has little utility if the primary interest is prediction of academic success.

The finding that the SOMPA correlates near unity with the WISC-R, and that mean scores only change slightly suggests that the additional exercise of questioning parents of children similar to those of this study about their financial status, living conditions and their own rearing yields questionable benefit in adding to the value of the WISC-R IQ alone. This does not imply, however, that the other benefits accruing from involving parents in the assessment process do not stand on their own. The above statement reflects that the emphasis of this study is on the predictive ability--future and concurrent--of intellectual ability tests.

With respect to the K-ABC, the fact that it is a new instrument, rather than just a renorming of an established instrument, brings with its use both promise and caution. The goal of the development of the K-ABC was to develop an intervention tool rather than a nosological indicator; that is, the emphasis was on prescription rather than labeling. While this was indeed the intent, the finding that the K-ABC MPC mean is almost identical to that of the full scale WISC-R IQ suggests that the K-ABC would be a suitable substitution for the WISC-R in student placement decisions. Much of the practitioner's knowledge and experience derived from working with Wechsler scales will translate into helping interpret scores derived from the K-ABC. One may expect an individual evaluated with the K-ABC to have overall skills very similar to those of an individual with similar global

scores evaluated using the WISC-R. Also, the Kaufman's lower age range relative to the WISC-R (K-ABC norms go to 2½) offers the opportunity to collect a greater sampling of the skills of low functioning students or young children. The absence of Age effects in any of the ANOVAs suggests that one gains this extra usefulness for younger and lower functioning subjects without compromising the test's usefulness for older students.

The similar WISC-R FS-IQ and K-ABC MPC means found in this study support Naglieri's (in press c) report of no significant discrepancy between Wechsler and Kaufman full scale scores on groups labeled as mentally retarded (mean WISC-R FS-IQ, 75.1; mean K-ABC MPC, 77.8), learning disabled (WISC-R, 96.8; K-ABC, 96.1) or "normal" (WISC-R, 108.3; K-ABC, 108.8). However, the present findings contrast to the findings of Kaufman and Kaufman (1983), who reported that K-ABC MPCs are generally lower than WISC-R FS-IQs, with a 3 to 4 point discrepancy for the normal population. This was presumed to be the result of using the ten-years-older WISC-R norms, which do not reflect the greater verbal ability and sophistication of children of recent years.

Kaufman and Kaufman report (and other findings concur) that K-ABC means are generally lower for normal and learning disabled children, and higher for mentally retarded children (mean difference, 6 points). Naglieri (in press a) found significant differences with the K-ABC MPC higher for a low functioning mentally retarded group. McCallum, Karnes and Edwards (1984) reported K-ABC MPCs to average 6 points lower for their sample of 41 gifted children in grades 4 - 6 in rural

Mississippi. Naglieri and Haddad (1984) also found a 5 point discrepancy (with MPC lower) in a sample of 33 learning disabled students.

There do not seem to be characteristics common to the studies cited above which would lead to suggestions regarding why this present sample's scores are similar to the findings of Naglieri (in press c) rather than the other studies cited. The high verbal ability loading of the WISC-R is cited by Kaufman and Kaufman (1983) as a source for the discrepancy between mean scores of the WISC-R and K-ABC for middle class subjects. However, the mean WISC-R Vocabulary subtest score (8.95, $N = 40$) and Similarities subtest score (8.50, $N = 40$) for this sample are higher than the overall WISC-R subtest mean (8.22, $N = 40$), so it could not be reasonably conjectured that poor verbal skills depressed the WISC-R FS-IQs to offset a true positive discrepancy to be found in the population at large.

The obtained correlation of .69 between WISC-R FS-IQ and K-ABC MPC is comparable to the correlation of .70 reported in the K-ABC manual (Kaufman & Kaufman, 1983) for a sample of 182 "normal children." For the same sample, Kaufman and Kaufman report a correlation of .76 between the WISC-R FS-IQ and K-ABC Achievement subtest. A comparable correlation of .82 was obtained from the present sample. Kaufman and Kaufman find it completely appropriate that the WISC-R should correlate more highly with the K-ABC's achievement battery than it does with the K-ABC MPC, as is the case with this sample. Because of the heavy weight given to verbal ability and factual knowledge, conventional IQs are seen as being measures of school-related accomplishments. For example, the WISC-R has an arithmetic subtest in the

Verbal scale--the K-ABC has an arithmetic subtest in its Achievement scale.

Finally, despite the absence of significant differences in obtained scores or accuracy in predictions, one should not assume that each of these three assessment methods can be interchanged casually. Making the SOMPA adjustments to the WISC-R FS-IQs made little difference for this sample, perhaps, but one cannot suppose that this low income, white, rural, Southern Appalachian sample with largely intact family units is typical of the population for which Mercer was attempting to program when developing the SOMPA. The sheer statistics of the SOMPA manipulations would have insured a much larger impact if the sample had been nonwhite or from a non-English speaking culture. This study was simply not a fair test of what the SOMPA was designed to do, and one cannot safely go further than stating that the SOMPA adjustments were marginally relevant, and their impact, trivial, for this population.

The WISC-R and K-ABC share about 50 percent of common variance--this amount appears sufficient to establish the K-ABC as a measure of what has been traditionally considered to be intelligence, and yet leave sufficient variance to establish its process-emphasis difference from the content-emphasis area staked out by the WISC-R. The evaluation of the K-ABC properly rests with factor analytic studies (e.g., will replication studies find the Sequential and Simultaneous scales loading into two factors without excessive overlap?) and, more importantly, with assessment of its utility in making prescriptive inferences. One has only to be reminded that psychologists have

written probably hundreds of journal articles trying to derive prescriptive implications from the subscale interactions of the WISC-R to appreciate the potential resident in the K-ABC with its scales expressly designed for such purposes.

The present study appears to offer a design which could be readily utilized with subject populations that would better test the assumptions of both the SOMPA and K-ABC. This study is representative of low income, white, rural eastern Tennessee children, and from the author's experience, representative of low income, white, rural southern Appalachian children. To more properly test both the SOMPA and K-ABC, samples drawn from nonwhite and nonmajority culture populations would be desired. Investigation with persons for whom English is the nonpreferred language, language impaired populations and handicapped students of various diagnostic categories would be desired for evaluating the K-ABC. Larger samples, with independent test administrations would allow greater sensitivity in rejecting null hypotheses. Furthermore, studies investigating the utility of the K-ABC in prescribing educational interventions are needed.

LIST OF REFERENCES

LIST OF REFERENCES

- Aliotti, N. C. (1980, July). Intelligence testing and the Larry P. decision: A trainer's perspective. The School Psychologist, p. 2.
- Anastasi, A. (1968). Psychological Testing (3rd ed.). London: Macmillan Company.
- Aronwitz, S. (1974). The trap of environmentalism. In A. Gartner, C. Greer & F. Riessman (Eds.), The new assault on equality: IQ and social stratification. New York: Harper and Row, 108-205.
- Bender, E. M. (1975). The Bender Gestalt Test for young children: Volume II, research and application, 1963-1973. New York: Grune and Stratton.
- Bernal, E. M., Jr. (1975). A response to "Educational uses of tests with disadvantaged subjects." American Psychologist, 30, 93-95. (Comment)
- Bernal, E. M., Jr. (1976). Perspectives on nondiscriminatory assessment. In T. Oakland (Ed.), With bias toward none: Non-biased assessment of minority group children. Lexington, KY: Coordinating Office for Regional Resource Centers.
- Bersoff, D. N. (1981). Testing and the law. American Psychologist, 36, 1047-1056.
- Bosma, B. (1973). The NEA testing moratorium. Journal of School Psychology, 4, 304-306.
- Brody, E. B., & Brody, N. (1976). Intelligence: Nature, determinants, and consequences. New York: Academic Press.
- Brown, F. (1979). The SOMPA: A system of measuring potential abilities? School Psychology Digest, 8, 37-46.
- Budoff, M. (1975). Measuring learning potential: An alternative to the traditional intelligence test. In G. R. Gredler (Ed.), Ethical and legal factors in the practice of school psychology. Philadelphia: Pennsylvania Department of Education.
- Carroll, J. B., & Horn, J. L. (1981). On the scientific basis of ability testing. American Psychologist, 36, 1012-1020.
- Cattell, R. B. (1968, March). Are IQ tests intelligent? Psychology Today, pp. 56-62.
- Chapanis, A., & Williams, W. C. (1945). Results of a mental survey with the Kuhlman-Anderson Intelligence Tests in Williamson County, Tennessee. Journal of Genetic Psychology, 67, 27-55.

- Clarizio, H. F. (1978). Nonbiased assessment of minority group children. Measurement and Evaluation in Guidance, 11, 106-113.
- Clarizio, H. F. (1979). In defense of the IQ test. School Psychology Digest, 8, 79-88.
- Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and white students in integrated colleges. Journal of Educational Measurement, 5, 115-124.
- Cleary, T. A., & Helton, T. L. (1971). An investigation of item bias. Education and Psychological Measurement, 28, 61-75.
- Cleary, T. A., Humphries, L. G., Kendrick, S. A., & Wesman, A. (1975). Educational uses of tests with disadvantaged students. American Psychologist, 30, 15-41.
- Cole, N. S. (1981). Bias in testing. American Psychologist, 36, 1067-1077.
- Cronbach, L. J. (1975). Five decades of public controversy over mental testing. American Psychologist, 30, 1-14.
- Council for the Society for the Psychological Study of Social Issues. (1969). SPSSI council statement on race and intelligence. Journal of Social Issues, 25, 1-3.
- Darlington, R. B. (1971). Another look at cultural fairness. Journal of Educational Measurement, 8, 71-82.
- Deutsch, M. & Brown, B. (1964). Social influences in negro-white intelligence differences. Journal of Social Issues, 20, 24-35.
- Diana v. California State Board of Education, C-70-37 RFP (N.D. Cal., 1973), stipulated settlement.
- Drew, C. J. (1973). Criterion-referenced and norm-referenced assessment of minority group children. Journal of School Psychology, 11, 323-329.
- Dunn, L. M., & Markwardt, F. C., Jr. (1970). Peabody Individual Achievement Test. Circle Pines, MN: American Guidance Service, Inc.
- Educational Testing Service (1979, Spring). Testing debate grows lively: Some new measures of ability expected. ETS Developments, pp. 4,5.
- Eells, K., Davis, A., Havighurst, R. J., Herrick, V. E., & Tyler, R. (1951). Intelligence and cultural differences. Chicago: University of Chicago Press.

- Eisman, B. S. (1958). Paired-associate learning, generalization and retention as a function of intelligence. American Journal of Mental Deficiency, 63, 481-489.
- Feuerstein, R. (1979). The dynamic assessment of retarded performers. Baltimore: University Park Press.
- Figueroa, R. A. (1979). The System of Multicultural Pluralistic Assessment. School Psychology Digest, 8, 28-36.
- Findley, W. Ability grouping. (1975). In G. R. Gredler (Ed.), Ethical and legal factors in the practice of school psychology. Philadelphia: Pennsylvania Department of Education.
- Fine, B. (1975). The stranglehold of the IQ. Garden City, NY: Doubleday and Company.
- Fishman, J. A., Deutsch, M., Kogan, L., North, R., & Whiteman, M. (1964). Guidelines for testing minority group children. Journal of Social Issues Supplement, 20, 129-145.
- Flaugher, R. L. (1978). The many definitions of bias. American Psychologist, 33, 671-679.
- Garcia, J. (1981). The logic and limits of mental aptitude testing. American Psychologist, 36, 1172-1180.
- Gartner, A., Greer, C., & Riessman, F. (Eds.). (1974). The new assault on equality: IQ and social stratification. New York: Harper and Row.
- Gay, G., & Abraham, R. D. Does the pot melt, boil or brew? (1973). Black children and white assessment procedures. Journal of School Psychology, 11, 330-340.
- Gerry, M. H. (1973). Cultural myopia: The need for a corrective lens. Journal of School Psychology, 11, 307-315.
- Gordon, E. W., & Terrell, M. D. (1981). The changed social context of testing. American Psychologist, 36, 1167-1171.
- Green, B. F. (1981). A primer of testing. American Psychologist, 36, 1001-1011.
- Griggs v. Duke Power Company, 401 U.S. 424 (1971)
- Grossman, H. J. (Ed.). (1973). Manual on terminology and classification in mental retardation (Special Publications No. 2). Washington, DC: American Association on Mental Deficiency.

- Guskin, S. L. Theoretical and empirical strategies for the study of the labeling of mentally retarded persons. (1978). In N. R. Ellis (Ed.), International review of research in mental retardation (Vol. 9). New York: Academic Press, 127-158.
- Haggard, E. A. (1954). Social-status and intelligence: An experimental study of certain cultural determinants of measured intelligence. Genetic Psychology Monographs, 49, 141-186.
- Haney, W. Validity, Vaudeville, and values: A short history of social concerns over standardized testing. (1981). American Psychologist, 36, 1021-1034.
- Hays, W. S. (1973). Statistics for the Social Sciences (2nd ed.). New York: Holt, Rinehart and Winston.
- Heber, R. A. (1961). A manual on terminology and classification in mental retardation. American Journal of Mental Deficiency (Monograph Supplement), 66.
- Herrnstein, R. IQ. (1971, September). The Atlantic Monthly, pp. 43-64.
- Hirsch, N. D. M. (1928). An experimental study of the east Kentucky mountaineers. Genetic Psychology Monographs, 3, 183-244.
- Hobson v. Hansen.
- Humphries, L. G. (1975). Addendum. American Psychologist, 30, 95-96.
- Isaacson, R. L. & van Hartesveldt, C. (1978). The biological basis of an ethic for mental retardation. In N. R. Ellis (Ed.), International review of research in mental retardation (Vol. 9), 159-183.
- Jackson, G. (1975). On the report of the ad hoc committee on educational use of tests with disadvantaged students: Another psychological view from the Association of Black Psychologists. American Psychologist, 30, 88-93.
- Jastak, J. (1949). A rigorous criterion of feeble-mindedness. Journal of Abnormal and Social Psychology, 44, 367-378.
- Jastak, J. F., & Jastak, S. (1978). The Wide Range Achievement Test (Rev. Ed.). Wilmington, DE: Jastak Associates, Inc.
- Jensen, A. R. (1968). Social class and verbal learning. In M. Deutsch, I. Katz & A. R. Jensen (Eds.), Social class, race, and psychological development. New York: Holt, Rinehart and Winston, 115-174.

- Jensen, A. R. (1969a). How much can we boost IQ and scholastic achievement? Harvard Educational Review, 39, 1-123.
- Jensen, A. R. (1969b). Reducing the heredity-environment uncertainty: A reply. Harvard Educational Review, 39, 449-483.
- Jensen, A. R. (1970). A theory of primary and secondary familial mental retardation. In N. R. Ellis (Ed.), International review of research in mental retardation (Vol. 4). New York: Academic Press, 33-105.
- Kagan, J. (1974). What is intelligence? In A. Gartner, C. Greer & F. Riessman (Eds.), The new assault on equality: IQ and social stratification. New York: Harper and Row, 114-130.
- Kaufman, A. S. (1979). WISC-R research: Implications for interpretation. School Psychology Digest, 8, 5-27.
- Kaufman, A. S., & Kaufman, N. L. (1983a). Kaufman Assessment Battery for Children: Administration and scoring manual. Circle Pines, MN: American Guidance Service.
- Kaufman, A. S., & Kaufman, N. L. (1983b). Kaufman Assessment Battery for Children: Interpretive manual. Circle Pines, MN: American Guidance Service.
- Kohlberg, L. (1968). Early education: A cognitive-developmental view. Child Development, 39, 1013-1062.
- Lambert, L. M. Exhibit C. IQ Trial. APA Monitor, pp. 8-9.
- Lambert, N., Windmiller, M., Cole, L., & Figueroa, R. (1975). Manual for the AAMD Adaptive Behavior Scale Public School Version, 1974 Revision. Washington, DC: American Association on Mental Deficiency.
- Laosa, L. M. (1976). Historical antecedents and current issues in nondiscriminatory assessment of children's abilities. In T. Oakland (Ed.), With bias toward none: Non-biased assessment of minority group children. Lexington, KY: Coordinating Office for Regional Resource Centers.
- Larry P. v. Riles, 495 F. Suppl. 926 (N.D. Cal. 1979)
- Lawler, J. M. (1978). IQ, heritability and racism. New York: International Publishers.
- Lora v. New York Board of Education, 456 F. Suppl. 1211 (E.D. N.Y. 1978)

- Mason, E. J. (1975). The teacher expectancy bias: What does it mean? In G. R. Gredler (Ed.), Ethical and legal factors in the practice of school psychology. Philadelphia: Pennsylvania Department of Education.
- Matarazzo, J. D. (1972). Wechsler's measurement and appraisal of adult intelligence (5th ed.). Baltimore: Williams and Wilkins.
- Maxwell, J. (1953). Social implications of the 1947 Scottish mental survey. London: University of London Press.
- McCallum, R. S., Karnes, F. A., & Edwards, R. P. (1984). The test of choice for assessment of gifted children: A comparison of the K-ABC, WISC-R, and Stanford-Binet. Journal of Psychoeducational Measurement, 2, 57-63.
- McNemar, Q. (1975). On so-called test bias. American Psychologist, 30, 848-851.
- Meeker, M., & Meeker, R. (1973). Strategies for assessing intellectual patterns in black, anglo and Mexican-American boys--or any other children--and implications for education. Journal of School Psychology, 11, 341-350.
- Mercer, J. R. (1973). Labeling the mentally retarded. Berkeley: University of California Press.
- Mercer, J. R. (1978). SOMPA conceptual and technical manual. New York: Psychological Corporation.
- Mercer, J. R. (1979a). In defense of racially and culturally non-discriminatory assessment. School Psychology Digest, 8, 89-115.
- Mercer, J. R. (1979b). Test "validity," "bias," and "fairness": An analysis from the perspective of the sociology of knowledge. Interchange, 9, 1-16.
- Naglieri, J. A. (in press a). Assessment of mentally retarded children with the Kaufman Assessment Battery for Children. American Journal of Mental Deficiency.
- Naglieri, J. A. (in press b). Concurrent and predictive validity of the Kaufman Assessment Battery for Children with a native American sample. Journal of School Psychology.
- Naglieri, J. A. (in press c). Use of the WISC-R and K-ABC with learning disabled, mentally retarded and normal children. Journal of School Psychology.

- Naglieri, J. A., & Haddad, F. A. (1984). Learning disabled children's performance on the Kaufman Assessment Battery for Children: A concurrent validity study. Journal of Psychoeducational Assessment, 2, 49-56.
- Oakland, T. (1973). Assessing minority group children: Challenges for school psychologists. Journal of School Psychology, 11, 294-303.
- Oakland, T. (1979). Research on the Adaptive Behavior Inventory for Children and the estimated learning potential. School Psychology Digest, 8, 63-70.
- Oakland, T. (1983). Concurrent and predictive validity estimates for the WISC-R IQs and ELPs by racial-ethnic and SES groups. School Psychology Digest, 57-61.
- Observer. (1970). Innate intelligence: Another genetic avatar. Psychological Record, 20, 123-130.
- Observer. (1973). The hereditarian manifesto. Psychological Record, 23, 417-422.
- Piaget, J. (1967). Six psychological studies (A. Tenzer, Trans., D. Elkind, Ed.). New York: Random House. (Original work published 1964)
- Prasse, D. P. (1981, June). School psychology and the law: The present and future. NASP Communique, pp. 3-4.
- Purvin, G. (1974). Intro to Herrnstein 101. In A. Gartner, C. Greer & F. Riessman (Eds.), The new assault on equality: IQ and social stratification. New York: Harper and Row, 131-162.
- Pyle, D. W. (1979). Intelligence: An introduction. London: Routledge & Kegan Paul.
- Rapier, J. L. (1962). Measured intelligence and the ability to learn. Acta Psychologica, 20, 1-17.
- Rapier, J. L. (1968). Learning abilities of normal and retarded children as a function of social class. Journal of Educational Psychology, 59, 102-110.
- Reed, E. W., & Reed, S. C. (1965). Mental retardation: A family study. Philadelphia: Saunders.
- Reschly, D. J. (1981). Psychological testing in educational classification and placement. American Psychologist, 36, 1094-1102.

- Riessman, F. (1974). The hidden IQ. In A. Gartner, C. Greer & F. Riessman (Eds.), The new assault on equality: IQ and social stratification. New York: Harper and Row, 206-223.
- Rohwer, W. D., Jr. (1968). Mental mnemonics in early learning. Teachers College Record, 70, 213-226.
- Rosenthal, R., & Jacobson, L. (1968). Pygmalion in the classroom. New York: Holt.
- Sabatino, D. A., Kelling, K., & Hayden, D. L. (1973). Special education and the culturally different child: Implications for assessment and intervention. Exceptional Children, 39, 563-567.
- Samuda, R. J. (1975). Psychological testing of American minorities: Issues and consequences. New York: Dodd, Mead and Company.
- Sarason, S. B., & Doris, J. (1969). Psychological problems in mental deficiency (4th Ed.). New York: Harper and Row.
- Schmidt, F. L., & Hunter, J. E. Racial and ethnic bias in psychological tests: Divergent implications of two definitions of test bias. American Psychologist, 29, 1-8.
- Schoenfeld, W. N. (1974). Notes on a bit of psychological nonsense: Race difference in intelligence. Psychological Record, 24, 17-32.
- Sharp, E. (1972). The IQ cult. New York: Copward, McCann and Geoghegan.
- Sherman, M., & Key, C. B. (1932). The intelligence of isolated mountain children. Child Development, 3, 279-290.
- Silverstein, A. B. (1970). The measurement of intelligence. In N. R. Ellis (Ed.), International review of research in mental retardation (Vol. 4). New York: Academic Press, 193-227.
- Silverstein, A. B. (1971). Deviation social quotients for the Vineland Social Maturity Scale. American Journal of Mental Deficiency, 76, 347-351.
- Simrall, D. (1947). Intelligence and the ability to learn. The Journal of Psychology, 23, 27-43.
- Spearman, C. E. (1904). "General intelligence" objectively determined and measured. American Journal of Psychology, 15, 206-221.
- Strauss, S. (1981). The gifted and non-gifted. Chicago: Saturn Press.
- Teachers and citizens protest the testing ripoff. (1980, January/February). NEA Reporter, pp. 3-4.

- Tennessee State Board of Education. (1979). Rules, Regulations, and Minimum Standards 1979-1980. Nashville: State Board of Education.
- Tennessee State Board of Education. (1982). Rules, Regulations, and Minimum Standards 1982-1983. Nashville: State Board of Education.
- Tennessee State Department of Education. (1982). Student Evaluation Manual. Nashville: State Department of Education.
- Tennessee State Department of Education. (1983). Administrative Policies and Procedures Manual. Nashville: State Department of Education.
- Terman, L. M., & Merrill, M. A. (1973). Stanford-Binet Intelligence Scale. Boston: Houghton Mifflin Company.
- Thorndike, R. L. (1971). Concepts of cultural fairness. Journal of Educational Measurement, 8, 63-70.
- Thorndike, R. L., & Hagen, E. (1977). Measurement and evaluation in psychology and education (4th ed.). New York: Wiley and Sons.
- Tilton, J. W. (1949). Intelligence test scores as indicative of ability to learn. Educational and Psychological Measurement, 9, 291-296.
- Wechsler, D. (1974). Manual for the Weschler Intelligence Scale for Children--Revised. New York: Psychological Corporation.
- Wesman, A. G. (1968). Intelligent testing. American Psychologist, 23, 267-274.
- Wheeler, L. R. (1932). The intelligence of east Tennessee mountain children. The Journal of Educational Psychology, 23, 351-370.
- Winer, B. J. (1971). Statistical principles in experimental design (2nd ed.). New York: McGraw-Hill.
- Wolman, B. J. (Ed.). (1973). Dictionary of behavioral science. New York: Van Nostrand Reinhold Company.
- Woodrow, H. (1946). The ability to learn. Psychological Review, 53, 147-158.
- Zeaman, D., & House, B. J. (1967). The relation of IQ and learning. In R. M. Gagne (Ed.), Learning and individual differences. New York: Charles E. Merrill Publishing Company, 192-207.

VITA

Stephen Leigh Tench was born and reared in Habersham County, Georgia. He was graduated from South Habersham High School in Cornelia, Georgia, in June, 1970. That fall he entered the Georgia Institute of Technology where he was supported by an Alumni Foundation National Merit Scholarship. After a two-year stint in a chemical engineering curriculum, he discovered the field of empirical psychology and was graduated with the Bachelor of Science in Applied Psychology degree in June, 1974.

He was admitted to the experimental psychology program in the Department of Psychology at the University of Tennessee, Knoxville, in the Fall of 1974. The following year he transferred to the school psychology program, and subsequently earned a Master of Arts degree in psychology in June, 1977.

He has been employed by the Grainger County (Tennessee) Board of Education since 1976, and has served as both school psychologist and coordinator of special education. He has been an occasional part-time private practitioner during those years, having obtained licensure as a psychological examiner. He was also appointed as an administrative hearing officer for the State Department of Education by the Tennessee Commissioner of Education, a position he currently fills.

He married Jean Elizabeth Marsee in March, 1978. They have one child, Sara Elizabeth, who was born in July, 1981.