Draft

AFFIDAVIT

I, Asa G. Hilliard, hold the Doctor of Education Degree in Educational Psychology from the University of Denver. I currently serve as Dean of the School of Education at San Francisco State University. I have had over twenty years of experience in teaching and administration in schools at all levels, from kindergarten to the university. My professional experiences have included close work with every economic level and with all the major ethnic groups. I have had firsthand teaching and counseling experiences with students who have been described as "slow learners," "disadvantaged," and "gifted." I have developed tests and assessment procedures for teachers and children in plural cultural settings. I have been involved in teaching and research on the subject of assessment of ethnic minority children, especially African-American children and youth. I have had extensive experience in direct observation and assessment of children in Africa over a period of six years. I have contributed to the professional literature in the area of assessment, and have presented professional papers on assessment in training workshops and at professional conferences.

I am qualified to evaluate written tests and examinations and other selection and/or screening procedures in terms of their racially discriminatory impact, their validity, their reliability, and especially their utility as pedagogical tools. I have made extensive studies of these tools as they apply to African-American children.

I have reviewed materials pertaining to the use of standardized tests in the Tucson School District Number One, Tucson, Arizona. I have reviewed these written materials with particular attention to the evidence for the impact of standardized testing on minority children. Based upon that review, I made the following observations which serve as the basis for my professional opinions to be presented later.

Standardized testing and other district assessment procedures are a major activity in the Tucson Public Schools. Intelligence (i.e., "IQ," "Ability," or "Aptitude"), Achievement, or Personality Tests or Assessments are in widespread use as a basis for making decisions about the educational experiences of all children. School district personnel assert the following things about these tests and procedures:

1. Some tests are said to be measures of "learning potential" for all students. District personnel apparently believe that they have an acceptable degree of predictive accuracy (i.e., the ability to <u>predict</u> accurately school success).

2. Standardized tests are used purportedly to determine the appropriate learning strategies for individual students or for groups of students who are classified as a result of scores earned on the test.

3. Some tests can be used to measure how well students have mastered the learning tasks which their teachers have taught.

4. Currently, utilized standardized tests can identify "emotionally disturbed" students amont all ethnic populations.

My review of the results of standardized testing and other generally applied assessment procedures in Tucson reveals that there is a consistant

and pervasive pattern of poor performance by Black, Brown and poor children. Poor performance is manifest particularly on tests of IQ, school achievement, and personality. Primarily as a consequence of these poor performances Black, Brown and poor children are in general:

1. Disproportionately over represented in classes for the mentally retarded and educationally handicapped.

2. Disproportionately under represented in classes for the gifted.

3. Disproportionately over represented in classes for the emotionally disturbed.

4. Disproportionately over represented in classes for communicative disorders, especially "speech defects."

Clearly then, the consequences of judgments and decisions which are based largely upon standardized assessment procedures are major ones for any child. Because of low scores a child may be labeled and forced to bear the associated stigma, and will most likely be formally or "de facto" tracked into a set of educational experiences or programs of study which will limit severely the opportunity for education and later work in life. In making any such judgments, therefore, the school district should, in my professional opinion, carry the burden of proof that tests are valid for the purposes stated and that their proper use will result in demonstrable benefits for students with whom they are applied. This principle of the burden of proof residing with the professional user has been reflected in the Uniform Guidelines for Employee Selection Procedures which have been proposed by the Equal Opportunity Coordinating Council, and published in the July 9th issue of the Federal Register. While the specific tests involved were tests for employee selection, the general principles are applicable to standardized tests which are used in any kind of screening.

> " These guidelines do not call for a user to conduct validity studies of selection procedures where no adverse impact results. ...The use of any selection procedure which has an adverse impact on the members of any racial ethnic or sex group with respect to hiring, promotion, or transfer, or other employment or membership opportunities will be considered to be discriminatory and inconsistent with these guidelines unless the procedure in validated in accordance with the principles contained in these guidelines or unless the use of the procedure is warranted under paragraph 3b. ...Accordingly, whenever a validity study is called for by these guidelines, the user should make a reasonable effort to investigate suitable alternative selection procedures which have as little adverse impact as possible for the purposes of determining the appropriateness of using or validating them in accord with these guidelines. If a user has made a reasonable effort to become aware of such alternative procedures in a validity study for a job or group of jobs, has been made in accord with these guidelines, the use of the selection procedure may continue until such time as it should be reasonably reviewed for its currency." (p.290)

Following these observations the EEOC has proposed some of the most stringent requirements for the demonstration of validity in order to provide equal protection for applicants of all racial, ethnic or sex groups. That <u>adverse impact</u> occurs as a consequence of labeling or identification associated with testing or other means of classification is clear to any sensitive observer. The United States Commission on Civil Rights asserts that Black children suffered a greater feeling of inferiority when, through attending a predominately White school,

The problem in Tucson is acute. The consequences of the application of standardized test results in scores for ethnic populations which are widely discrepant from the White middle class regular pattern of performance. The source of the measured differences in performance among groups is due either to the ability, personality, and achievement of students, to errors in the assessment process, or inequities in the instructional process. In such a condition specific professional action is required on the part of responsible educators.

> "With full knowledge of the segregated consequences of the procedures school authorities continue to catagorize and separate. To perform an act with known consequences is to intend the consequences; this concept runs through the law. In the context of testing and tracking, the supposed dichotomy between de facto and de jure segregation is simply not important. Where school officials have for some time employed academic standards for testing devices in the classification and assignments of pupils, and when the result has been a continuing and consistently disproportionate impact upon minority groups, discriminatory intent can be inferred from the natural, probable, and foreseeable effect of perpetuating racially identifiable classrooms. It is the school authorities who are imposing a segregated educational structure and it is they who are mandated to change. As one commentator put it, "in every case of racially imbalanced school sufficient responsibility can be ascribed to government to satisfy the requirement that stems from the equal protection clauses post scription of unequal treatment by government." Indeed, unless we are to read Swann one way for previously de jure systems in the South and another for equally segregated schools in the North, we are obligated to find that racial imbalance is itself a denial of equal protection. Given the affirmative duty imposed in Swann, any school board which does not take all feasible steps to alleviate or eliminate classroom imbalance can be held responsible for its continuation." (Sorgen, 1973, p.165-166)

Sorgen here has summarized a number of legal rulings. In short, the Courts have apparently confirmed what every professional educator should, in my opinion, accept that evidence of inequity and outcome requires an extra professional effort to remedy the situation.

Apparently, the official position of the professional staff of the Tucson Public Schools is that the IQ test differences and, therefore, other test differences, which are found among many ethnic groups in Tucson are real differences in ability, achievement, and personality between minority and poor populations and the middle class White population which is assessed. In a document which came in 1969 from the Superintendent and his staff (document reference 3916, Ruben Salter). The position seems to be taken that not only were the wide ethnic ability and differences among groups in Tucson to be expected, in fact, there should be expected even higher numbers of low performers than are found at present. Apparently, aligning itself with "the President's Committee on Mental Retardation," the district seemed to accept the idea that "low income," "economin stagnation," "malnutrition," and "poor health" "while not directly linked, more than a coincidence is obviously at work."

(Note: In the affidavit, please insert here the facts that are cited in Ruben's document, page 3916 on page 7)

they are "accorded separate treatment with others of their race, in a way which is obvious to them as they travel through the school to their classes." (Sorgen, 1973, p. 1131) When this adverse impact is present, in view of the gravity of the consequences, special effort is required to insure the accuracy and competence of assessments.

In the final analysis, any dedicated educator must ask, in my opinion, that professional assessment demonstrate consequences which have a substantial educational purpose and which produce a demonstrable benefit to the student. Unfortunately, research on the application of standardized testing, particularly with minority populations, fails to reveal that there is a demonstrable benefit for these children.

> "Educational research poses a quite different kind of challange to present classification practice. It has increasingly undermined one of the central premises of sorting: that it benefits students. The research concerning the educational effects of ability grouping and special education, reveals that classification as it is typically employed, does not promote individualized student learning, permit more effective teaching to groups of students of relatively similar ability, or indeed accomplish any of the things it is ostensibly meant to do. Educational efficacy studies generally find no effects, or marginal adverse effects on achievement and attitude for students who are classified. When these students are compared with non-grouped peers. These findings applied both for the average and the slow student (in some studies, the brightest students appear to benefit slightly from the grouping) and in evaluation both of ability grouping in special education programs for the mildly handicapped in England and other European countries as well as in the United States.
>
> The research indicates that the classification effectively separates students along racial and social class lines, and that such segregation causes educational injury to minority groups. It also suggests adverse classifications stigmatize students, reducing both their self-image and their worth in the eyes of others. Indeed, if the researchers had their way the profession would now be 'writing an epitaph for grouping.'
>
> Even those who accept the basic premises of school sorting have reason to question whether schools can adequately do the job. Two re-tests of students assigned to classes for the retarded revealed notable system made errors. In Washington, D.C. the school system itself conducted the re-testing; it found that 2/3 of the students placed in special classes in fact belonged in the regular program. A study of 378 educable mentally retarded students from 36 school districts in the Philadelphia area indicated 'the diagnosis for 25% of the youngsters found in classes for the retarded may be considered erroneous. An additional 43% may be questioned."' (Kirp, 1973, p.714)

The research cited by Kirp above is a typical pattern in urban schools throughout the Nation. There are many other studies in the professional literature which show similar results. There is little in the literature to demonstrate a link between standardized assessment and successful instruction.

> "Despite these problems, the consistency of results among all the studies (in particularly among those most carefully executed) is impressive. It indicates that most school classifications have marginal and sometimes adverse impact on both student achievement and psychological development." (Kirp, 1973, p.717)

It is interesting to note that according to Sorgen the President's Committee on Mental Retardation has also joined in calling for a moratorium on the use of tests in identifying and classifying school children until a more appropriate instrument might be developed.

No materials or documents have been made available up to this point in this case that district expectations for minority and poor children's performance has changed in any substantial way. In fact, other statements from the Superintendent seem to show a consistent district posture that is essentially in agreement with the historical results of testing. The Superintendent has been quoted as saying the following:

> "We believe that this charge is based upon an erroneous assumption that language skills in themselves constitute the reason for the relatively low achievement rate of the children in question. We believe that it is abundantly clear that the socio-economic level of the neighborhood and the family is the culprit. The identical phenomena is evident in all- Anglo neighborhoods. Our teachers who are familiar with schools in Mexico report that children in those schools who come from low socio-economic neighborhoods do not achieve as well as children in more affluent neighborhoods.
>
> Language then is not the major factor that makes a difference in achievement. The Mondale Report as well as local research amply demonstrates that academic achievement of children is directly related to the economic background of the homes. It is not a racial, ethnic or language problem. Multimillions of dollars have been poured into 'compensatory' education programs across the Nation with very little success. We can only conclude that schools are being blamed when the real fault lies with society's own isolation of people into economic strata (Refrence document 15853, Ruben Salter). On still another occasion the Superintendent was reported to have spoken to the Board of Education on the matter of student mobility and scores '... the influx of floaters' -- people who drift from City to City -- and the children have school problems as a result." (Document 15172, Ruben Salter)

The importance of these pronouncements for an examination of standardized testing and its use in the Tucson schools is that they seem to suggest a kind of expectancy at the highest levels of school leadership. That the low performance of students is caused by factors over which the school has no control such as poverty, family background, student mobility, nutrition and so forth. The Court, in Larry P. vs Wilson Riles, shifted the burden of proof to defendants and refused to assume that there was any relationship between race and ability to learn unless proof could be provided to that effect. The school district, in its arguments, suggested that poverty, malnutrition and mental retardation were related, however, without proof the court assumed that the ability to learn would be the same in all groups in the population. In other words the apparent correlation between poverty and scores on IQ tests is not the same as a real correlation between poverty and aptitude. Then too, the danger of the existence of negative expectations is that they appear to become a self fulfilling prophecy. There is impressive research (Rosenthal and Jacobsen) (Brophy) (Rist) to show that the expectations of school people have a way of influencing the student response to assessment and thereby becoming the self fulfilling prophecy. If this is likely to be the case, and certainly the expectancy literature is well known among professional circles today, one would certainly expect that minimum standards of quality control of every assessment would be developed and applied in order to insure accuracy in an area so

vital to the future of children who are to be served by the school. One would expect both accurate, precision diagnosis (testing) and evidence of an unassailable match among testing, instructional strategies and student gains. In my opinion this has not been the case in Tucson.

I will now illustrate by reference to practices in IQ testing, achievement testing and personality testing. If testing and assessment are not to be regarded as discriminatory I would expect to find two vital things. 1) valid instruments and procedures for all students including minorities. 2) a pattern of use which leads clearly to student gains as a consequence of what is learned from the testing. In other words, both the instruments (tests) and the use to which the instruments are put would be subject to the most rigorous requirements for the demonstration of validity. In view of challanges to the validity of tests for minority populations and in view of professional knowledge (or assertions of) the relationship between expectancy and pupil performance, a kind of due process would require the rigorous quality control for all professional tools as indicated above.

A person who is engaged in research may be justified in conducting all kinds of experiments and trial runs, however, professional practitioners are not researchers and are responsible for using only those practices which are known to be beneficial to students. In the case of standardized testing it is not the case that something is better than nothing in the testing of students if students do not get better when professional practice is applied.

IQ Test in the Tucson School District Number One

The following individual and group IQ (aptitude, intelligence, or ability) standardized test are among those in use in the school district according to school district documents:

1. Stanford Binet
2. Wechsler Adult Intelligence Scale (WAIS)
3. Wechsler Intelligence Scale for Children (WISC)
4. Wechsler Pre-school Scale of Intelligence (WPSSI)
5. Lorge-Thorndike
6. Otis-Lennon

It is appropriate, at this point, to share an observation about test validity from the proposed EEOC Guidelines on Employee Selection Procedures published in the Federal Register of July 14, 1976 and alluded to above.

> "Under no circumstances will be the general reputation of a selection procedure, its author or its publisher or casual reports of its validity be accepted in lieu of evidence of validity. Specifically ruled out: are assumptions of validity based on a procedures name or descriptive labels; all forms of promotional literature; date of bearing on the frequency of a procedures usage; testimonial statements and credentials of sellers, users or consultants; and other non-empirical or anecdotal accounts of selection practices or selection outcome... an employment agency is not relieved of its obligation herein because the user did not request such validation or has requested the use of some lesser standard of validation than is provided in these guidelines." (p. 291)

There is no evidence that any of the standardized tests which are used to diagnose specific learning difficulties in the Tucson School District Number One are used in a way that lets the teacher know what to do next in teaching practice.

The real use of these instruments is admittedly to help school people to predict- to guess better at what the student will be able to do sometime in the future, and in general. In practice they tell almose mothing new about a child that would not be guessed from a combination of school grades, poverty and ethnic group membership; they simply confirm the judgments which could be made on these bases while relieving school personnel of the responsibility, for a large part, of decision making. However, the question for schools should be "can good teaching change things for children." If so, then the tests do not measure intelligence or ability at all, only whether the student and the examiner have learned the same things. In fact, while there is some relationship among scores which a person would earn on the tests above, the definitions of intelligence are, in fact, different for each of the tests listed. For example, in the Buros Mental Measurements Yearbook the standard reference for providing professional reviews and evaluations of existing standardized tests, the Stanford-Binet Intelligence Scale, an individual test which is listed as one which has been used in the Tucson District, was reviewed by Dr. David Freides, Associate Professor of Psychology, Emory University, Atlanta, Ca. Among Dr. Freides' comments are the following:

> "Today we remain a long way from systematic knowledge of the nature of intelligence and hence we still do not know exactly what to measure, but it is clear this is the direction to go which may explain why the Wechsler scales hace largely superseded the Binet." (p. 773)

In another review of another test, the Lorge-Thorndike also listed as one of those in use by Tucson schools, Carol K. Tittle, Assistant Professor of Teacher Education at City University of New York has stated the following:

> "In their study of the comparability of IQ's obtained from the Lorge-Thorndike and four other intelligence tests administered to pupils in Iowa in grades 4, 7, and 10, they note that the correlations were quite variable and, in most cases, below the reliabilities of the test indicating that the tests were measured somewhat different in traits. The non-verbal IQ scores of the L-T had lower correlations with the other intelligence test scores than the verbal. The Hieronymus and Stroud Study also provides comparability data on the IQ's derived from the 1954 and 1963 editions of the Lorge- Thorndike. In their study the newer edition yielded slightly lower IQ's for the verbal scores in grades 4, 7, and 10 and for the non-verbal grades 7 and 10."

I could go on with comparisons among other commonly used "tests of intelligence," however, the main point is that there is no uniform definition of the construct of intelligence not is there uniformity among instruments which purport to measure intelligence. Consequently, the criteria for selection from among the available instruments to measure intelligence are essentially determined by local or individual psychologists decisions. Perhaps even more important, however, than the lack of a degree of congruence of agreement among the different definitions and measures of intelligence is the fact that throughout the reviews and in the bibliographies associated with the reviews of these tests, there are no studies which demonstrate the utility of intelligence testing for the development of instructional strategy with minority children. A student of low measured ability does get assigned a "place" very frequently as a consequence of a low score. The assumption implicit in the definition of "ability test" is that potential is a fixed thing. Good teaching, therefore, should not affect the outcome on a test. Therefore, one way to check the validity of a test is to see if teaching can make a difference in the scores earned.

If a school district has not lost faith in the capacity of its children, it could and should do something on its own to determine if its assessments were, in fact, accurate, unbiased, true measures of minority student ability. Specific professional practices would and should be designed to remediate difficulties with testing rather than simply to accept them, especially when adverse impact exists or the lack of demonstrable benefit obtains. Therefore, professional staff would, in my opinion, be expected in the interests of children to be actively in persuit of, or in search for, validation for or equitable alternatives to presently used tests. This principle is reflected in the Equal Employment Opportunity Coordinating Council Guidelines which were referred to and which were published in the Federal Register Volume 41, No. 136 on page 29017.

> "Accordingly, whenever a validity study is called for by these guidelines the user should make a reasonable effort to investigate suitable alternative selection procedures which have as little adverse impact as possible for the purpose of determining the appropriateness of using or validating them in accordance with these guidelines...an employment agency is not relieved of its obligation herein because the user did not request such validation or has requested the use of some lesser standard of validation than is provided in these guidelines. The use of an employer's agency does not relieve an employer or labor organization of its responsibilities under federal law to provide equal opportunity or its obligations as a user under these guidelines."

There are no consumer protection agencies to act on behalf of clients where standardized testing is concerned. Clients cannot simply accept the word of test publishers, who have a vested interest in improved earnings, for that publishers opinion about the quality of the instruments. No one would submit to medical treatment under those circumstances. No one would even purchase a used car without a demonstration that the instrument worked. Certainly students have a right to expect demonstrable benefit before being subjected to professional procedures which are really in the research stage rather than in the application stage.

A district seeking alternatives might include such activities as the following: (or equivalent activities which a staff competent in assessment could easily design) For a district under Court Order where attention has been focused for such a long period of time, on this specific question one would expect an aggressive drive to insure a practice of non-biased assessment. For example, a district could do some of the following things to check the claims of unregulated test manufactures:

1. The Tucson District could conduct a small pilot project to determine if populations which have been assessed as mentally retarded are truly unable to learn beyond the predictions which are made based upon scores earned on standardized tests of intelligence. This would mean utilizing the services of teacher known to be skilled in cross-cultural assessment and cross-cultural teaching. These teachers would teach the regular school subjects. The progress of students could then be compared to the predictions which were made by those who had assessed them as low in the first place. There is impressive evidence to demonstrate that, particularly in the case of minority children, dramatic gains are frequently made when excellent teaching is provided. For example, Project SEED (Special Educational Enrichment for the Disadvantaged), which is housed at the Lawrence Hall of Science in Berkeley, Ca., now a nationally funded project, has been able to demonstrate that minority and poor children learn Mathematics at and even above the level of their middle class counterparts. In fact, minority and

poor children are taught college level Mathematics in addition to Arithmetic. These projects have been conducted throughout the length and bredth of the Nation and in foreign countries. They have been evaluated by the Northwest Regional Laboratory for Educational Research and Development and by the Math Department at Cal. Tech., as well as by many other interested investigators. Project SEED, in all probability, is the most thoroughly evaluated educational intervention about which I have information. Without any special equipment and in classes of approximately 30, SEED trained teachers are uniformly successful in working with populations which are thought to be unable to learn as predicted by intelligence tests. One would think that such an intervention would have great appeal in a school district where minority students do not achieve.

2. The school district could do a systematic search for culturally relevant and valid assessment devices.

3. The school district could do a review of literature on the validity of the assessment instruments which they use particularly as the research bears upon the applicability of those instruments to minority population. In other words, it is a professional obligation to document the validity of the instruments for use with the populations which are now being served. For example, in the Buros Mental Measurements Yearbook a review was given by Alvin G. Burstein, Professor and Chief of the Division of Psychology at the University of Texas Medical School at San Antonio, Texas. He was reviewing the Wechsler Adult Intelligence Scale and his comments about that instrument are as follows:

> "It is in that sense unfortunate that work relating socio-cultural factors and WAIS performance is so scanty and of such poor quality, the tendency to use overcrude independent variables (e.g., Black vs. White without regard to cultural characteristics) or to sample inadequately is one common defect; and there is a tendency to design the studies as though variation from group to group in the configuration of scores were the primary interest rather than variation in the predictive meaning of scores. It is only in this latter sense that the issue of "cultural bias" has meaning. From the point of view of rhetoric and semantics, substituting the term "culturally specific validities" the term "cultural bias" might help to sharpen scientific issue of the cultural specificity of the predictive validities of WAIS scores has simply not been dealt with adequately in the literature." (Page 787)

In the face of such information, it seems ludicrous to continue to apply the Wechsler Adult Intelligence Scale to minority children. In the Buros review of the WPPSI, the Wechsler test for young children, A. Oldridge and E. Allison make the following observations:

> "The reliability of the verbal performance and full scale IQ's are satisfactory but the sub-test are not

sufficiently stable to be of much value for individual use."

Macnemar's (1956, page 127) statement regarding the WAIS is also relevant to the WPPSI.

> "The author of the WAIS has attempted an impossible task. The construction of a scale to measure general (global) intelligence which at the same time will provide differences among sub-tests which are of diagnostic value. Only the global objective appears to be well achieved." (page 809)

It is interesting to note in the light of the quotation above that users of individual tests of intelligence claim that there is more accuracy in individual than in group tests of intelligence. In fact, for minority children and perhaps even for majority children, the diagnostic and prescriptive value of these instruments for instructional purposes has yet to be demonstrated. In fact, the Stanford- Binet appears to one reviewer to be so poor that he made the following concluding observation in his review.

> "The Stanford- Binet Intelligence Scale is an old, old vehicle. It has lead a distinguished life as a pioneer in the bootstrap operation that is the assessment enterprise. Its time is just about over, Kequiesceat in pace." (David Friedes, page 426)

The Otis-Lennon Mental Ability Test is among those listed as being available for use in Tucson. As with all tests of "mental ability" there are serious difficulties, particluallry when they are applied to populations for whom they were not specifically designed. Buros Seventh Mental Measurement Yearbook carries three reviews on the Otis-Lennon. Two of the reviews are reasonably favorable. However, it is interesting to note that even though the first reviews are generally favorable, while the third is somewhat critical, in neither of the three reviews does the Otis-Lennon get high marks for validity, and nothing is said about the application of the test to minority children. Dr. John Milholland, Professor of Psychology at the University of Michigan, Ann Arbor, has noted that the rationale for the Otis-Lennon has come from "Vernon's description of the structure of mental abilities, embodying two major divisions, "verbal- educational" and "practical-mechanical," integrated into Spearman's g. The Otis-Lennon Test aims to cover only the verbal-educational half of the structure." (p. 690) The terms "verbal-educational" and "practical-mechanical" are really defined in terms of performance on the Otis-Lennon test. The literature on the Otis-Lennon does not lead us to independent observation of children which would enable us to observe "verbal-educational" behavior. This is typical of the situation with IQ testing in general. It is the kind of condition which lead one outstanding psychologist to observe that "intelligence is what intelligence tests measure." The reasoning is circular. The consequences of permitting test makers to be both judges and jury of validity for their own commercial product, especially where minority children are concerned, leaves the children at the mercy of people who seldom claim to have cultural sophistication covering the varity of groups which make up the American mosaic, and who have economic as well as professional interests in the continuing large scale listing.

Dr. Milholland also makes the point that I made earlier, "the question arises however, to what purpose is all this effort expended? Just what

is the value of national norms? Has one learned anything by discovering that a school pupil, a grade, or a system is this or that far above a nationwide norm figure? Does it help to guide instructional policy?" (page 690) In none of the three reviews is their any discussion of studies which have demonstrated the instructional utility for the Otis-Lennon,nor is there any discussion of claims by the test publisher that such utility exists, except in the very broadest sense. For example, we find in the second review by Dr. Arthur Smith the statement, "The Otis-Lennon test can be a useful tool for teachers, administrators, and counselors." However, there is no indication as to what this usefulness is. Nowhere in any of the studies cited in Buros or in the "validity studies" which are reported on the Otis-Lennon is there any thing to suggest what a teacher, counselor, or administrator is to do with the results, other than the traditional labeling or classification which comes as a consequence of "predictions." At a very general level, the first reviewer, Dr. Milholland, has stated, "The Otis-Lennon Test should perform well the functions it is intended to serve. These explicitly do not include measuring innate learning potential, and in the manual, special caution is advised in interpreting results for children who do not have normal backgrounds and motivation." (page 370) While this caveat is offered, neither the Otis-Lennon Test Manual not the reviewers comment offer any guidance for interpreting results for children who do not have "normal backgrounds and motivation." Presumably, "normal" backgrounds and motivations refer to children who have backgrounds and motivation similar to that of the group upon which the test was standardized.

In the review by Dr. Arden Grotelueschen, a closer look at the validity question was taken. Dr. Grotelueschen has observed:

> "In contrast, some sections of the technical data and developmental research parts of the manual lack specific information essential to test consumers. The most serious lack of data is that of validity. The authors intend well by stating that validity studies are being conducted and that the results will appear in a forthcoming technical handbook. Information on the procedures for standardization, scaling and norming is presented in detail. Moreover, the procedures for each appear to have been carefully conceptualized and conducted." (page 692)

This is a typical situation with tests of mental ability. Information on standardization, scaling, and norming is written with a precision which can easily cause the reader to overlook the fact the necessary validity studies are either absent or inconclusive. So here we have an excellent example of an instrument which has not completely developed, which is still in the research stages, which was developed without any particular reference to minority cultural groups, and which is listed by the School District as being available to be applied to students. In my professional opinion, this instrument would be inadequate for use with White middle-class students. To use it with the variety of minority groups which are present in a urban environment such as Tucson brings professional practice to something more like guess work than science. Without the careful study of groups to which the tests are to be applied, and without the development of a specific

cultural expertise by those who are to use the test, I firmly believe that the practitioner will be doing the minority children a grave injustice.

Dr. Thomas Hilliard, a clinical psychologist in San Francisco who has served as an expert witness in many court cases involving the assessment or misassessment of Black clients by assessors who are unfamiliar with the cultural background of their clients, has observed:

> "It has been my experience that most White psychologists are unaware of the severe problems and limitations of traditional psychological instruments concerning standardization, validity, and reliability with Blacks and other ethnic minorities, and problems in their administration, scoring and interpretation. Nor are they aware of the empirical research of clinical data that indicates that the race of the examiner greatly affects the examinee's responses in many cases. For many Blacks, this impact is negative. Finally, these psychologists often are not community oriented and are totally unaware of community mental health and other resources .... In fact these severe limitations in clinical training and awareness of current thinking in Black psychology and mental health raises serious questions as to whether many white clinicians can make unbiased and accurate assessments of minority clients."

In Grotelueschen's review, in an incredible statement, he gives a perfect example of the state of the art in mental ability testing.

> "The authors should be applauded for facing up to an embarassing situation, that of making explicit a rational for a test which has been in existance for around 50 years. In the development of the rationale it was reasoned that the Otis-Lennon series should continue to be a broadly based measure of general mental ability defined more specifically as verbal-educational g. It would have been desirable if the authors had presented factor analytic evidence to compliment the logical claims for the test. The omission of a factor study of available data is of concern to this reviewer, especially since coorelational evidence interpreted by the authors to support the construct validity of the test is inconclusive. For example the authors interpret the high correlations between the Otis-Lennon and various mental ability batteries as evidence for construct validity. No explanation is given however to account for similar high correlations between the Otis-Lennon and various achievement test scores. (The Otis-Lennon is not unique in this respect) The high correlation, for example, between the Otis-Lennon and the Iowa Tests of Basic Skills composite score leaves only 5% of the non error variance unexplained at the fifth and eighth grade levels, when both predictor and criterion are corrected for attenuation. This finding is not appreciably different from that observed between forms J and K of the

> Otis-Lennon itself. Thus the Otis-Lennon gives ample evidence for predicting scholastic success. However, without stonger evidence for construct validity, it may be concluded that the predictability of scholastic success is due to the fact that the Otis-Lennon is a direct measure of scholastic success. Stated more practically, present evidence indicates that schools with an achievement testing program may not be adding anything unique to the total testing program by using the Otis-Lennon as a measure of general ability." (pages 371-2)

In the quotation above the following things should be noted.

1. The author points out that this test has been used for 50 years without an explicit rationale.

2. Many of the claims around test validity are based on "logical" reasoning while observations of the actual behavior of children are missing.

3. Even the relatively cheap, factor analytic statistical method has not been used.

4. This well known test of "mental ability" adds almost nothing in the way of new knowledge about a child which would not have been learned by reference to existing tests of "achievement".

5. There is nothing said about instructional utility.

Given the above, a minority child might be given the Otis-Lennon and information from that test might be used to assist in making placements into a class for the mentally retarded or the gifted. This is like doing surgery with a rusty nail. The most that any professional could hope for from such an instrument is to justify a decision that had already been made by other means. This instrument, as is the case with others which are used to measure "mental ability" for minority children is for these children a discriminatory device.

4. The Tucson District could select an assessment staff to include those who, in addition to the possession of ordinary assessment skills, are intimately familiar with the culture and background of the minority population which are served. They should also have demonstrated the ability to make valid assessments with minorities.

5. The Tucson District could conduct systematic training and assessment for any existing staff who administer and interpret tests in order to insure that they are able to apply their skills competently with minority population. For example, their results might be compared with those which are obtained by assessors who are known to be skilled in the assessment of minority populations.

6. The Tucson School System could keep meticulous records of test results and recommended instructional practice to determine if the "prescriptions" which are based upon test scores result in "appropriate learning" for the children.

7. The Tucson School System could conduct an aggressive program of consultation with minority professionals who are skilled in assessment of specific minority populations, especially when these professionals are not available in the district. Such professionals can help to

varify diagnoses and to train existing staff, however, an appropriate quality program of assessment would be one in which adequate ethnic representation is present in order to do the quality job demanded.

8. The Tucson District could review every existing test and evaluate them by the professional criteria which have been established for the evaluation of tests and diagnostic techniques by the American Psychological Association and by the American Educational Research Association. Especially important in this regard is that the tests which are used meet the most stringent, predictive content and construct validity, and that these be demonstrated for each ethnic group which is served. General figures on validity and reliability for so called "norm" or normal populations are totally inadequate for application to the diverse minority populations which make up many urban areas.

9. The Tucson District could demand that sales people or test producers who market assessment devices provide demonstrations of the utility of any device which they select to apply to their pupil populations.

There is no evidence that any serious attempt has been made to guarentee the validity and the reliability of the assessment process in Tucson. The fact that tests correlate with each other is clearly insufficient evidence of validity. The fact that tests may correlate with future performance is also insufficient evidence of validity. The school is not in the horse betting business, but the teaching business. The professionals job is not simply to make guesses about the future but to figure out the problems of the present. There simply is, therefore, no evidence in the documents made available to me to show that adequate quality control of testing and assessment process using such activities as those described above, or any other set of systematic attempts to assure accuracy and equity. Some of the activities in the progress report show that "in-service" sessions were "attended". Just as students are not given IQ achievement scores or adequate grades for "attendance" the quality-control of the assessment process would dictate that those who go through training but also be required to demonstrate minimum competencies in cross-cultural assessment as a consequence of that training.

It is my professional opinion that the existing practices in IQ testing consistently and unfairly discriminate against Black, Brown and poor children resulting in a systematic underestimation of their ability.

1. Unlike the White middle class child, most minority children must have their ability tested in an unfamiliar language or dialect, denying to that child the use of his or her lifetime store of experience and language.

2. Unlike the White middle class child, the minority child frequently must have his or her ability tested using vocabulary which is not commonly used in his or her environment or community.

3. Unlike the White middle class child, the minority child frequently must be tested by professional psychometrists or teachers who have no real understanding of or appreciation for or vested interest in his or her cultural patterns and with whom the child may be unable to identify and accept as a helper. This affects the degree to which a fundamental essential in any assessment rapport can be presumed to be present.

4. Unlike the White middle class child, the minority child frequently is expected to be a poor performer by the very people who are responsible for his or her assessment.

It can be seen, therefore, that to continue with "standardized" procedures, when those procedures are stacked in favor of a particular ethnic group, is to operate with a built-in disadvantage for other ethnic groups. It is, therefore, a discriminatory practice to begin a comparative assessment of ability by not giving all participants the same advantage of a comparable assessment environment. To fail to do this is to require Black, Brown and poor children to compete in leg irons.

In short, there is no evidence of any instructional utility for the mass of IQ testing which is done in Tucson beyond the administrative justification for sorting or placement. For minority students, these tests have virtually no utility, are discriminatory and result, for many, in labeling and stigmatization which is clearly unnecessary, unfair and professionally unjustifiable.

Frequently, users of standardized tests ask if IQ tests are not used, what will we put in their place. The question carries an assumption. The assumption is that there is some utility to the tests which are being used. My comments are intended to indicate that not only are tests inherently biased, but that in addition, even if they were not they serve no useful purpose where minority children are concerned.

Achievement Test in the Tucson District Number One

District officials indicate that a wide variety of tests of achievement are available for use. Among them are the following:

1. Boehm Test of Basic Concepts

2. Metropolitan Achievement Test

3. Stanford Early School Achievement Test

4. Stanford Reading Test

The problem of discrimination against minority students in the use of achievement tests is quite different from the problems associated with the testing of intelligence. Whatever "intelligence" is it is not definition something over which the school is supposed to have control. On the other hand, at least some of what a student has "achieved" comes as a consequence of teaching. Therefore, it is entirely appropriate for school personnel to try to assess the consequences of school experiences by use of standardized tests of achievement provided that certain basic conditions are met.

1. The test selected must be a reliable instrument. It must be a consistent instrument, in other words.

2. The test selected must be a test that matches the things which are taught in a school district. In other words, it must have content validity.

3. If an achievement test is supposed to be a diagnostic instrument as well, the things measured on the test must really be the components or sequences which are required for student learning.

If these conditions are met, no child, minority or otherwise, has any cause to resist assessment on the grounds of being subjected to a technically inadequate instrument. On the other hand, if the test measures one set of things which is different from the things which are taught in school then it would be most inappropriate to use that test in school assessment. If minorities are found systematically to be low on instruments which are technically inadequate, they have every right to call the school system to task.

My review on this matter reveals that no information was provided to prove the match between the test content and the content of courses which are taught to students. This is a serious matter. The differential performance of students on achievement tests demands that if equity is to be assured, there be a demonstration of an accurate match between the content which is assessed through standardized testing and the teaching which is done in the Tucson schools.

However, this is not the most serious matter. The real case for discrimination in the use of achievement tests is less a technical matter than a matter that the selection of any test is also a de facto commitment to the educational goals for a district. A district indicates by its selection of tests and by the use of those tests, and in communicating about student gains in terms of those tests that it accepts certain content as important. This means that the key decisions about what content is important most frequently been made in an arena which is free from representative government, where there is little or no input from minority people. Such a condition leaves minority people in particular at the mercy of those who construct test item. As a result there is a degree to which a school system's curriculum resources will be devoted to teaching content which at best ignores important educational priorities and at worse frequently presents a false distorted and demeaning picture of the role of minorities in our society. Therefore, when test content is determined external to a district by professional people who are not responsible for the content of the test to any public body, minority people suffer still another blow from the very institution to which so many look as their main hope for extrication from a cycle of poverty and discrimination.

A school district does have the option and the professional responsibility to conduct locally a review of a selected test and to determine its applicability to the local school situation. To be non-discriminatory this determination should be done in cooperation with representatives of the constituency groups which are served by the school. I found no evidence in the available documents which describe the testing practices of the district to indicate that any systematic review of achievement test content for ethnic acceptability and equity had been conducted.

In the case of "diagnostic" achievement tests, a special potential for discrimination can be cited as an example. The Stanford Reading Test is listed as one of the tests which is used in assessing the achievement of students in Tucson. The Stanford Achievement Test has the following sub-test in skill areas:

1. <u>Sub-test</u>

   Dictated vocabulary
   Word reading
   Paragraph reading
   Word study skills
   Listening comprehension

2. <u>Skill areas</u>

   Reading and literature
   Non-fiction and reference

Skill areas continued

Math and Science
Social science
Word reading
Explicit meaning
Implicit meaning
Consonant clusters
Consonant sounds
Consonant diagraphs
Long vowel sounds
Short vowel sounds
Variant vowel sounds
Influence classification
Global meaning
Explicit detail
Implicit detail
Conceptual meaning
Logical analysis

It takes only a cursory review to see that many of the things listed under reading skills are things which many adults, who are known to be good readers would be incapable of achieving. For example, the matter of vowel sounds, clusters is not a metter of standardized practice. Across the length and breadth of our nation there are varieties in pronunciation and articulation, and yet, if a portion of a reading test score is made on the basis of articulation, then the population on which the test was standardized has a distinct advantage over all other populations. Further, it is confusing to mix the way one pronounces and hears words with scores of "reading" which most non-professionals infer to mean reading comprehension. It took many people a long time to understand, for example, that a Southern accent for White, middle class people, was not evidence of inferior English, but simply a different style of pronunciation.

Tests of Personality

The tests of personality which are in use by the district will not be listed here. Suffice it to say that in no case of personality test currently in existence has sufficient work been done taking into account the high level of cultural variation, both between and within cultural groups to provide a reliable and valid index of personality development, especially when assessed cross-culturally. The extent to which such instruments or tests of personality development, especially when assessed cross-culturally. The extent to which such instruments to tests of personality are used by culturally unsophisticated examiners as a means of placing children in classes such as for the emotionally disturbed, represents the exact degree to which discrimination can be said to occur.

Summary

My review of the materials from the Tucson School District Number One leads me to the following conclusions.

1. The pattern of use of standardized tests, particularly standardized tests of aptitude, IQ, or ability, has led to adverse impact on Black and Mexican-American children in the Tucson schools.

2. In the face of this adverse impact, no substantial educational purpose has been demonstrated for IQ tests which are used with minority children.

3. No substantial educational benefit has been shown for minority children as a consequence of the use of IQ tests.

4. There is no demonstrable utility which can be shown to lead to educational gains when these tests are applied to minority children.

5. In the face of these difficulties, there still is no systematic data to show that any of the IQ tests in use by the district are designed in such a way that they take into account the unique cultural experiences of minority children.

6. There is no systematic data to show that the Tucson School District has instituted the quality control of the assessment process which would assure the elimination of misassessment.

7. Within the broad area of quality control, there is no evidence that an attempt has been made to verify the competence of those who are responsible for administering, scoring, and interpreting tests with minority populations.

8. There is no evidence of an agressive quality program for in-service training of professionals who are responsible for the assessment of minority children to insure the quality of their work with these children.

9. There is evidence that the official position which has been expressed by the Superintendent of the District is that a low level of expectation exists for minority children, justified because of the childrens poverty, nutrition, etc.

10. In the case of standardized tests of achievement, while many which are in use appear to have from minimal to acceptable reliability, there is no evidence that (in the face of adverse impact of testing on minorities) that a rigorous and aggressive review of existing measures of achievement have been conducted to insure the following three things.

    a) that there is in fact a close match between the content which is taught in the schools and that which is assessed on the standardized test.

    b) that the content which is taught and tested for is ethnically fair.

    c) that the content which is taught and tested for reflects priorities appropriate to minority populations.

11. In the case of personality testing, no attempt has been made here to do an assessment of the particular tests in uses in the schools. If culturally inappropriate, standardized tests of personality are even more grossly unfair. Quite simply, the standardized tests must meet the professional standards for use with minority populations just as they are expected to do with the so called "normal" population.

In the light of the conditions which have been summarized above, it is my professional opinion that the pattern of use of standardized testing with minority children in the Tucson City School District is such that it constitutes gross, continuing, negative, and unfair discrimination against those children. In the case of aptitude and personality testing, the state of the art is such that existing standardized tests are experimental devices. They are unproven, and yet they assist educational decision makers in arriving at educational de-

cisions which will affect the life chances of minority children. The situation with achievement testing, while more hopeful, leaves much to be desired.