Alne, Dennis James

EVIDENCE OF BIAS IN THE WAIS: A COMPARISON OF BLACKS AND WHITES

City University of New York

PH.D. 1984

University Microfilms International 300 N. Zeeb Road, Ann Arbor, MI 48106

EVIDENCE OF BIAS IN THE WAIS:

A COMPARISON OF BLACKS AND WHITES

By

DENNIS J. ALNE

A dissertation submitted to the Graduate Faculty in Educational Psychology in partial fulfillment of the requirement for the degree of Doctor of Philosophy, The City University of New York.

1984

This manuscript has been read and accepted for the Graduate Faculty in Educational Psychology in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

June 21, 1984
date

David Rindskopf
Chairman of Examining Committee

June 25, 1984
date

Shirley Feldmann
Executive Officer

Professor David Rindskopf

Professor Alan Gross

Professor Barry Zimmerman
Supervisory Committee

The City University of New York

Abstract

EVIDENCE OF BIAS IN THE WAIS:

A COMPARISON OF BLACKS AND WHITES

by

Dennis J. Alne

Adviser: Professor David Rindskopf

Recently there has been increasing concern over the use of individually administered intelligence tests for making educational decisions regarding minority groups. It has been contended that these tests are biased against blacks (Williams, 1971), and their use in the evaluation of black students has led to underestimating actual intelligence. Due to this underestimation, inappropriate educational decisions are being made (Williams, 1974).

Beginning in the late 1970s researchers began to pay serious attention to investigating such contentions. Previous studies of bias in the Wechsler scales, have been on younger children. For this, and other reasons, the WAIS was evaluated in the present study.

Data was collected on 358 black and 256 white high school students. The students had been referred to school psychologists for an evaluation of learning and or behavior problems. All evaluations took place between January 1980, and June 1982.

Consistent with the literature, the definition of test bias followed in the present study is that a test is biased if two individuals of equal ability, but of different group membership, do not have equal probability of obtaining the same test scores. Based

on this definition several hypotheses regarding test bias were investigated.

While the groups were found to differ significantly in factor structure, some measure of differences were expected due to uncontrollable factors, Further, the difference between the white and black samples were not nearly as great as the differences in factor structure between the present white sample and a normal white sample. In addition, the direction of item bias did not favor one group significantly more often than the other when the usual internal criterion of ability was used to match groups.

In view of the findings it was concluded that the WAIS has comparable internal validity for the two samples, but is a better measure of a global factor of intelligence for whites in this study. This last finding is supportive of the contention that more blacks than whites in the present sample may be misdiagnosed as mildly retarded, and reinforces the importance of including other data when making this diagnosis in minorities.

Acknowledgements

This dissertation would not have been written were it not for the support and encouragement of those closest to me. Therefore, this dissertation is dedicated to my parents, Clara and Olaf Alne, and to my wife Doreen. I also wish to thank David Rindskopf for all the help and guidance he gave me along the way.

TABLE OF CONTENTS

TABLE OF CONTENTS (continued)

TABLE OF CONTENTS (continued)

List of Tables

List of Tables (continued)

List of Figures

Evidence of Bias in the WAIS:
A Comparison of Blacks and Whites

Introduction and Statement of Purpose

In recent years there has been increasing concern over the use of individually administered intelligence tests for making educational decisions regarding minority students. This concern has involved not only psychologists and educators, but public interest groups and legislators.

Individually administered intelligence tests, the most frequently given being the Wechsler scales, are major components of the evaluation process by which students are assigned to special education classes. There have been a number of court cases challenging the legality of using these tests for making decisions about placing minority students in such programs, especially in classes for the educable mentally retarded.

The most famous court case is that of Larry P. (Larry P. et al. v. Wilson Riles et al., 1974, 1979). Plaintiffs in this litigation contended that intelligence tests are racially and culturally biased against blacks, thereby denying them their rights of equal protection under the law in violation of their fourteenth amendment rights. Another class action suit, brought on behalf of bilingual Mexican-American children (Diana vs. California State Board of Education, 1973) charged that intelligence tests are "prejudicial to these children in regard to their native language, cultural background and normative standardization."

The result of these and other court cases has been a restriction of the use of the tests, both by the federal government, and by individual states, most notably California. Legislation at the federal level has had an impact on assessment procedures. For example, Public Law 94-142, The Education For All Handicapped Children Act of 1975, mandates that assessment instruments be selected and administered so as not to be racially or culturally discriminatory. The rules and regulations for implementing this law (Federal Register, 1977) require that the instruments be validated both for the groups they are to be administered to, and for the specific purpose for which they are to be used.

The most outspoken critics of individually administered intelligence tests have argued that the tests are biased against minority groups because the content of the items and or the skills needed to answer them correctly are more familiar to members of the majority white culture (Boozer, 1978; Dove, 1968; Hardy, Welcher, Mellits and Kagan, 1976; Hunt, 1974; Mercer and Lewis, 1978; Sowell, 1977; Williams, 1970, 1971, 1974, 1975; Zaref and Williams, 1980). In a study by Sattler and Kuncik (1976), it was found that practicing psychologists themselves believe individually administered intelligence tests are less valid for minorities and therefore are not as good a measure of their actual intellectual ability, often underestimating it.

For the most part, those mentioned as critics of intelligence tests feel that the types of experiences they have been exposed to places minorities at a disadvantage when answering intelligence test items.

To support their contentions, such critics have pointed out items that are "obviously" biased to favor members of the majority white culture. An example that is often given is an item from the Wechsler Intelligence Scale for Children-Revised (WISC-R) Comprehensive subtest. This item asks what you should do if a boy (girl) much smaller than yourself starts to fight with you. The correct answer is to ignore them. Critics contend that minorities have learned from experience to fight back as a means of survival. Therefore, they claim the item is biased. Based on such arguments, Williams (1970), who is past president of the Association of Black Psychologists of the American Psychological Association, called for a ban on the intelligence testing of minorities. Williams (1971) asserts that intelligence tests are underestimating the intellectual ability of blacks and leading to incorrect educational decisions concerning them. According to Williams (1971), blacks are being misclassified as mildly retarded and put in special education classes, and are being counseled away from college when they have the ability to be successful at it.

An assumption that has been made by those critics who offer arguments based on bias in the content of test items is that it is possible to infer the differential effect that certain item content will have on the responding to items by various racial groups.

Studies by Plake (1980) and Sandoval and Miile (1980) force one to question all arguments which have been based on this assumption. For example, Sandoval and Miile had 100 undergraduates of three races attempt to pick out those WISC-R items which, due to their content, they believed should be more difficult for minorities. The authors then correlated the ratings of these subjects (based on a five point

scale) with actually computed item difficulties taken from the SOMPA standardization group (Mercer and Lewis, 1978). Sandoval and Miile found that there was no correlation between the subjective ratings and objectively computed item difficulties. In addition, the item from the Comprehension subtest mentioned earlier as thought to be biased against minorities was actually found to be slightly less difficult for them in this large representative sample.

Due to such findings, and to the nature of the issues involved (potential mislabeling of minority students) it becomes imperative to carry out studies in which we objectively evaluate the "bias" inherent in a test so one can avoid making statements that may prove to be misleading and unwarranted.

Whether or not a test is as good a measure of ability in one group as it is in another can be investigated in either of two ways. An external investigation would involve determining if the relationship of test scores and an objective outside criterion was the same in both groups. In the case of intelligence tests this has largely been accomplished by regressing intelligence test scores on achievement test scores (Conroy and Plant, 1965; Covin and Covin, 1976; Dean, 1979; Reschly and Sabers, 1979; Reynolds and Gutkin, 1980; Reynolds and Hartlage, 1979).

The other method is to examine certain internal characteristics of test scores to see if the factorial structure of the test is the same for both groups and if groups are responding similarly to the individual test items (Jensen, 1980). This method was used in the present study. Specifically, the purpose of the present investigation

is to investigate the presence of "bias" in an individually administered test of intelligence; the Wechsler Adult Intelligence Scale (WAIS).

A test is defined as biased when two individuals of equal ability, but of different group membership, do not have equal probability of obtaining the same expected test scores. From this definition various hypotheses can be derived. Two individuals of equal ability, but of different group membership, will not have equal probabilities of obtaining the same expected test scores if the test is not measuring the same abilities in both groups or if the individual test items are not equally measuring the specific abilities they were designed to measure (Shepard, Camilli and Averill, 1981).

Test bias can be investigated by utilizing the linear structural relations methodology of Joreskog (1969, 1970) to analyze the subtest covariance matrix. Item bias has received considerable attention in recent years; a number of item bias detection procedures have been proposed (Shepard, Camilli and Averill, 1981). One of these, log linear models, will be used in the present study.

The findings of the present investigation will enable one to evaluate the contentions of intelligence test critics (Williams, etc.) as regards the Wechsler Adult Intelligence Scale (WAIS). That is, are the inferences being drawn concerning intellectual ability equally appropriate for blacks and whites?

The WAIS was chosen for the present study because it has received very little attention as regards black-white comparisons when compared to the Wechsler Intelligence Scales for Children (WISC/WISC-R). For example, while a number of investigators have examined the factorial structure of the WISC and WISC-R for blacks and whites (Gutkin and

Reynolds, 1981; Reschly, 1978; Semler and Iscoe, 1966) there are no such studies of the WAIS. Further, all studies of item bias in the Wechsler scales have been confined to the WISC (Miele, 1979) or WISC-R (Cotter and Berk, 1981; Oakland and Feigenbaum, 1979; Sandovaal, 1979).

The voluminous body of research studies involving the WAIS is mostly clinical in nature. The WAIS has been investigated for its usefullness in the identification of learning disabilities, organic brain syndromes, schizophrenia, criminal tendencies etc. Few studies have directly compared racial or ethnic groups and those involve few subjects. The present investigation will provide initial, and much needed, evidence as to the appropriateness of generalizing the inferences drawn from test scores and past research findings to different racial groups.

Literature Review

In 1904, the Minister of Public Instruction for the Paris school system appointed a commission to study the identification and education of intellectually subnormal children attending the Paris schools. Alfred Binet became a member of that commission, and in collaboration with William Simon, prepared the first Binet-Simon Scale (Anastasi, 1968). The scale, an individually administered test, was designed to cover a wide variety of functions, which Binet regarded as essential components of intelligence. The purpose of this and future revisions of the original scale was to enable one to identify those children who could not benefit from the traditional curriculum due to limited intelligence. This initial purpose is still one of the major reasons that individually administered intelligence tests are administered to school children today.

Since the introduction of the Binet-Simon Scales and their successor the Stanford Binet (Terman and Merrill, 1960) there have been many measures of intelligence developed. None of these measures has achieved the widespread acceptance of the Wechsler Intelligence Scales (Wechsler, 1958).

The Wechsler scales have been used extensively, both in the United States and in a number of foreign countries. While their primary use remains in the evaluation of intellectual ability, they have also been used considerably for research purposes.

During the 1960s and 1970s these tests came under increasing attack from individuals and groups who felt that tests were under-predicting the actual intellectual ability of minority groups, most

notably of blacks, and therefore inappropriate decisions were being made about them (Williams, 1970). Class action suits were brought against State Boards of Education (Larry P. et al. v. Wilson Riles et al., 1974, 1979; Diana et al. vs. California State Board of Education, 1973) and resulted in legal rulings that placed restrictions on the use of these tests for minorities. Among other things, it was the opinion of the courts that the content of many of the test items favored members of the majority white culture, hence placing minorities at a disadvantage on the tests.

Basis of Anti-Test Attitude

As research was compiled showing whites to have significantly higher IQ scores than blacks (averaging approximately one standard deviation) on the WISC and WISC-R (Covin and Hatch, 1976, 1977; Holowinsky and Pascale, 1972; Kaufman and Doppelt, 1976; Vance, Hankins and McGee, 1979) and on the WAIS (McGrevy, Knouse and Thompson, 1974; Overall and Levin, 1978) a growing number of psychologists and educators questioned the validity of these tests for minorities (Boozer, 1978; Hardy, Welcher, Mellitts and Kagan, 1976; McShane, 1980; Mercer, 1973; Smith, Hayes and Solway, 1977; Williams, 1970, 1971, 1974, 1975). Sattler and Kuncik (1976) found that practicing psychologists themselves questioned the validity of individually administered intelligence tests for minority groups. These authors asked 110 psychologists to estimate the "true IQ's" or "effective intelligence" of black, white and Mexican-American subjects from an evaluation of WISC protocol patterns. They found that minority IQ's were estimated to be higher than whites with the same protocol patterns. To help understand these findings, Sattler and Kuncik did some further

investigating and concluded that minority IQ's were estimated to be higher than whites because this large sample of psychologists generally believed that the WISC was less valid for minorities, and, in fact, often underestimated their actual intelligence.

An assumption implicit in many of the anti-testing arguments is that there are no group differences in genotypic intelligence. What differences are found are assumed to be due to differential accessability to the kinds of learning experiences required to be successful on intelligence test items. Hunt (1974) states that while one's genetic endowment sets limits on intellectual potential there is wide variation within these limits. This variation is largely due to the types of environmental experiences one has had. Therefore IQ tests may not serve as accurate indications of hereditary capacity or intellectual potential.

A study by Scarr (1978) is pertinent in this regard. Scarr noted that many studies (reviewed by Jensen, 1969) have shown that individual differences in IQ within the middle class white population are more related to genetic than to environmental differences. Scarr questioned whether this finding held for disadvantaged groups. She collected data (IQ scores) on both advantaged and disadvantaged identical and fraternal twins in Philadelphia. Scarr found that genetic differences accounted for about half of the IQ differences among middle class children (consistent with other studies), but very little of the IQ differences among lower class children. In lower class groups, identical twins were no more alike than were fraternal twins. She concludes that many disadvantaged children do not have the kind of

home and neighborhood environments that give them the skills required for IQ tests. They may develop other skills, however, school type vocabulary and reasoning skills are poorly developed.

Probably the most outspoken critic of individually administered intelligence tests is Williams. As past president of the Association of Black Psychologists of the American Psychological Association he has argued that the intelligence and academic achievement of black children has been misunderstood (Williams, 1970). According to test critics such as Williams, individually administered intelligence tests are biased because they are differentially valid for white and black groups. Williams asserts that due to the varying experiences they have beeen exposed to, the tests are not equally measuring the construct of intelligence in both groups.

To date, the arguments upon which such assertions have been made lie mostly in the content of the individual test items. For example, Williams asks: "Is it more important to know George Washington's birthday or Booker T. Washington's birthday; Malcolm X's last name or the author of Hamlet?" In William's opinion those who have considered IQ and intelligence to be one and the same are mistaken. An IQ score is relative and contains considerable errors of measurement, whereas intelligence indicates the extent to which an individual is able to understand and adapt to his or her environment. "The content of these tests is biased to favor whites and has led to the mislabeling of black children" (Williams, 1971). Williams contends that the intellectual ability of blacks is being underestimated by these tests and hence inappropriate decisions are being made. He states that

blacks are being misclassified as mildly retarded and assigned to special education classes, and further are being counseled away from college when they actually have the intellectual ability to be successful at it.

In an effort to show that blacks would score higher than whites on a test with items loaded to favor members of the black culture, Williams (1975) constructed the Black Intelligence Test of Cultural Homogeneity (BITCH). While it is true that blacks score higher than whites on this test, Long and Anthony (1974) have shown that black children will do poorly no matter how culture specific the test. In addition, Matarazzo and Wiens (1977) administered both the BITCH and the WAIS to black and white police applicants. They found that while there was considerable overlap of WAIS scores, there was almost no correlation between the two tests, thus raising the question as to whether the BITCH has any validity as a test of intelligence. The argument of Matarazzo and Wiens appears to be as follows: The WAIS is a widely accepted measure of intelligence. Even if critics claim that the WAIS underestimates black intelligence they usually do not reject the fact that it measures intelligence. Therefore, the failure of the BITCH and an accepted measure of intelligence to correlate forces one to question whether the BITCH measures intelligence.

Another study that attempted to show how alterations of test content would effect black and white subjects is one by Schmeisser and Fergusen (1978). These authors employed an experimental design to help determine whether the type of test content influenced group performance. Schmeisser and Fergusen chose reading passages for the English Usage and Social Studies Reading Tests of the American College

Testing Program that were considered to be more reflective of either the black or white culture. For example, two of the passages of the Social Studies Reading Test dealt with the history of blacks in America. The third passage was considered to favor the white culture and dealt with seventeenth century English history. The tests were then given to a representative sample of black and white subjects. The authors concluded that test content did not have a major effect on any of the variables studied.

Even if Schmeisser and Fergusen were to have found that test content did not have a major impact on group performance, another problem with the arguments of test critics emerges. Most of those who consider the WISC/WISC-R and WAIS to be biased base such assertions upon subjective judgments of bias in the content of the items which favors members of the majority white culture. To these individuals it seems that many of the test items must be more difficult for blacks because of their content. One such example was mentioned in the introduction.

The studies of Sandovaal and Miile (1980) and Plake (1980) noted earlier in which it was found that there was no correlation between subjective judgments of item difficulty and objectively computed item difficulty force one to question all such assumptions.

Due to such findings, and to the importance of the issues involved (potential mislabeling of minority students) it becomes imperative to carry out studies in which we evaluate the "bias" inherent in these tests so as to avoid making statements concerning individuals which may prove to be misleading and unwarranted.

Up to now the term "test bias" has been used with no real attempt made to define its meaning. Before proceeding we will explore this multifaceted concept and attempt to delineate some of the many issues involved.

Definition of Test Bias

A number of definitions of test bias have been proposed. None is perfectly applicable in all situations nor satisfactory for all audiences. The difficulty in delineating one definition for every occasion arises from the complex interactions between how a test measures what it is supposed to measure and whether the intended content is proper for a particular use (Shepard, Camilli and Averill, 1981). In effect, a continuum can be said to exist which ranges from situations in which the test is a valid measure of some particular skill or construct in both groups, and is being used appropriately for the particular purpose for which it was built, to the other end of the continuum in which the test does not measure what it is purported to measure in both groups and is being used for a purpose very different from that for which it was intended. A test may be biased when used for a particular purpose or with a particular group, and yet that same test may be perfectly unbiased when used for a different purpose or with a different group. Due to the complex issues involved in establishing one acceptable definition of bias many researchers conclude that a totally unbiased test may never exist. Therefore, a number of models of fair selection procedures have been developed. Basically, these procedures attempt to minimize the unfairness against one group when a biased test is used in the selection process. Gross and Su (1975) state that each of the models implicitly assigns loss values (percentage of incorrect decisions) and sets acceptance scores for each group that minimize the expected loss values.

The present study will focus on "bias" that is due to certain internal characteristics of the test itself. For our purpose, test bias can best be viewed as a type of invalidity where the inferences drawn from test scores are differentially valid for two groups. This notion can best be explained by way of example.

On a test of numerical word problems, if the vocabulary of the items is more familiar to one group than to another then the test may be said to be placing one group at a disadvantage, and any inferences made from the test scores may be differentially valid for the groups concerned. As an extreme example, items asking sixth grade boys and girls to compute the batting averages of star ballplayers would favor the boys. Therefore, the items are not equally measuring the skill (computing averages) that they are supposed to be measuring, and the test scores may result in invalid inferences being drawn; that is, girls are weaker on certain math skills.

According to Shepard, Camilli and Averill (1981), tests that are built to measure psychological constructs require major inferences, for example, that the test items tap the intended construct. If these inferences are less sound for one group than for another, differential validity (bias) will result. Green (1975) used the example of a measure of word knowledge taken as a measure of intelligence. Green states that although the test may accurately assess individual differences in reasoning ability for affluent whites, it more likely measures vocabulary or opportunity to learn for blacks. Inferences concerning the intellectual ability of the two groups which are based on the test scores are likely to be differentially valid, and may result in inappropriate decisions being made regarding blacks.

The definition of test bias to be followed in the present investigation is similar to the one proposed by Reynolds (1982): "Test bias denotes constant or systematic error, as opposed to chance or random error, in the estimation of some value; for our purposes here, this constant or systematic error is usually due to group membership..., and occurs in the estimation of a score on a psychological or educational test..."

The definition of test bias to be followed in the present study is a somewhat more specific version of this and can be inferred directly from the arguments of the critics of the Wechsler Intelligence Scales. The WAIS will be said to be biased if two individuals of equal ability, but of different group membership, do not have equal probability of obtaining the same expected test scores.

While this definition may appear largely theoretical in nature it will allow the development of testable hypotheses which can help evaluate the charges of bias in the WAIS made by critics such as Williams.

Now that test bias has been defined, ways in which it can be investigated will be discussed.

Evaluation of Test Bias

The schemes that have been most frequently recommended for the evaluation of test bias are based largely on the traditional trinary conceptualization of validity: content, criterion related and construct (Berk, 1982). For example, Jensen (1980) defines the criteria of test bias as falling under two main headings: (1) predictive validity criteria and (2) construct validity criteria. Predictive validity criteria refers to how well scores on the test enable one to predict scores on an external criterion. Construct validity criteria can be considered under two main categories: External and internal.

External construct validity refers to the correlations of test scores with other variables independent of the test itself. Thus, a test's predictive validity may also enhance its construct validity. What is important here is the pattern of relationships found among the test and various external variables. On theoretical grounds, test scores should be correlated with certain variables, and should be uncorrelated with other variables.

Internal construct validity refers to various quantifiable features of the test data themselves, such as item statistics,[1] as well as the factorial structure of the test. In this regard, a test is not biased against one group if the same factor structure is present in both groups (Reynolds, 1982), and the individual test items equally tap the constructs that they were designed to measure (Shepard, Camilli and Averill, 1981).

A word of caution must be included at this point; a compartmentalized conceptualization of bias (bias due to differential

content validity, differential predictive validity etc.) could foster the notion that there are equivalent alternative approaches to its investigation, and a test maker need only pick one of them. This could lead to incorrect conclusions; for example, that because there is no evidence of content bias, the test is completely unbiased in relation to all test uses. In addition, the three categories are inadequate for classifying the diverse forms of bias that may be suspected in terms of possible test score interpretations, uses, and inferences.

The recent trend toward a more unified conceptualization of validity seems to suggest a more useful framework for evaluating test bias. Questions of validity are seen as questions of what may be properly inferred from a test score; validity refers to the appropriateness of inferences drawn from test scores (American Psychological Association, 1974). Messick (1979) argues that we are not well served by labeling different aspects of a general concept with the name of the concept, as in criterion related validity, content validity, or construct validity, or by proliferating a host of specialized validity modifers, such as discriminant validity, trait validity..., etc. The substantive points associated with each of these terms are important ones, but their distinctiveness is blunted by calling them all "validity." Worse still, any one of these so-called validities, or a small set of them might be treated as the whole of validity, while the entire collection to date might still not exhaust the essence of the whole.

This conceptualization of validity would de-emphasize the types of validity and instead emphasize the data gathering process and the

methods to obtain the kinds of evidence essential to answer the charges of bias that have been brought against a test. The test maker would concentrate on the inferences to be made with the scores and the sources of bias that could invalidate those inferences. Subsequently, evidence would be accumulated to ascertain the presence or absence of bias.

Consistent with this conceptualization, the present study will attempt to gather evidence that can be viewed as either supporting or refuting the charges of bias that have been made against the WAIS.[2]

It has been charged that, because of the different experiences to which they have been exposed, the same factors or psychological constructs are not equally being measured in black and white groups, and that therefore, the WAIS is not accurately measuring the intellectual ability of blacks (Williams, 1971), resulting in inappropriate educational decisions being made about them.

These charges can be investigated in two ways: Predictive validity studies would enable one to determine if the test scores could predict academic success (high school or college) equally as well in both groups. Internal validity studies would enable one to determine the extent to which the same factors were being measured by the test in both groups, and if there are individual test items that are being responded to differentially by the two groups.

Predictive Validity Studies

While the present investigation will focus on what Jensen has termed internal validity evidence of test bias, it would be remiss not to make a point about external or predictive validity studies. Showing that the academic achievement of blacks and whites could be predicted equally by WAIS scores would greatly help to answer the charges of bias that have been directed at the test. Unfortunately such evidence is lacking.

Anastasi (1968), writing thirteen years after the introduction of the WAIS, deplored the lack of validity data present at that time. Anastasi implored researchers to carry out systematic investigations of the test's validity as this would help to strengthen the interpretation of test scores.

A review of the research literature concerning the WAIS leads to the conclusion that the request made by Anastasi has largely been ignored. Not only are there very few studies of the WAIS predictive validity, but only three studies specifically comparing black and white groups could be found. All of these studies regressed WAIS scores on the scores of other intelligence tests. Their similar purpose is typified by this statement by one of these investigators (Peteroy, 1980): "This was done in order to establish the concurrent validity of the Quick Test for both blacks and whites." Peteroy's reasoning assumes that the WAIS has been shown to be a valid measure of the intellectual ability of blacks. Unfortunately, this has not been shown to be true except in roundabout ways. The children's version (WISC/WISC-R) have been shown to have acceptable predictive validity for blacks. Since the content of these tests is similar it is assumed

that the WAIS must also have acceptable predictive validity for blacks. Until such evidence is obtained, any generalizations of research findings to blacks should be qualified.

This review of internal validity will include studies of the WISC and WISC-R, since past research has utilized these scales. Because the two forms of the Wechsler Intelligence Scales (adult scale and children's scales) are highly similar in format and based on the same hierarchical theory of intelligence, reviewing the research findings from the children's scales should help one to develop hypotheses concerning expected findings on the present study of the adult scale.

Internal Validity Studies

Internal validity studies are seen as providing evidence concerning test bias (Jensen, 1980). The factorial structure of the test scores should be highly similar for both groups (Reynolds, 1982) and the individual tests should be measuring equally the specific factors or skills which they were designed to measure (Shepard, Camilli and Averill, 1981).

If these aspects of test scores (factor structure and characteristics of responding to test items) are dissimiliar for two groups then we have evidence that the test is not measuring, to the same extent, the same thing in both groups. Therefore, the inferences made from the test scores may well be differentially valid for the groups concerned.

The investigation of the internal validity of the WAIS for blacks and whites consists of two stages. In the first, the subtest variance-covariance matrices are analyzed to determine if the factorial pattern is invariant over both groups. Following this, in stage two, the individual test items are analyzed to investigate the presence of differential patterns of responding by groups.

Each stage will now be discussed in greater detail. The relevant literature will be explored. The statistical methods to be employed will be detailed, and finally, based on past findings, the hypotheses for the present analysis will be delineated.

Stage One

Factor analytic studies of the WAIS. The initial factor analytic studies of the WAIS appeared to find great discrepancies in the number of factors present as well as in the pattern of factor loadings. For

example, Cohen (1957) found three major factors and two that were specific to certain subtests. He defined the major factors as Verbal Comprehension (factor 1), Perceptual Organization (factor 2) and Freedom from Distractability (factor 3). These three factor names have been widely accepted, and are still used today. The one exception to this being the third factor (Freedom from Distractability) which is sometimes called Memory. Another investigator, Saunders (1959), on the other hand, found nine factors, and the Vocabulary subtest was not even administered to his subjects. While these two studies represent extremes as far as discrepant number of factors found, Nerviano (1974) noted that trying to summarize the literature was impossible because of the variations in the factor analytic techniques employed, data manipulations involved, and rotations utilized. Nerviano decided to reanalyze the data from the factor analytic studies of the WAIS making use of the same factor analytic technique. He felt that this would allow him to more effectively summarize the past research findings.

Nerviano obtained the original data from thirty studies (some were reanalyses of the same data). A total of twenty one different, usable, WAIS intercorrelation or unrotated factor matrices provided the data for his reanalysis. While Nerviano actually employed three different methods of analysis and three types of rotations, he notes that the findings were highly consistent. The principal parameters regarding the acceptance of any factorial solution were a joint consideration of the absolute value and proportion of variance for each successive variate, as well as the replicability, interpretability, and cross-methodological stability of that particular solution. Because of slightly greater overall clarity, results from the principal

components analysis along with direct oblimin rotation were chosen for data presentation. The subjects included in these studies represent a great variety of ages (16 - 75+), of psychopathologies (normal - psychiatric inpatients) and cerebral impairments (brain injured - epileptics).

Some relevant results from Nerviano (1974) will now be discussed. Because the subjects in the present study range in age from 16 - 20 (few 20 year olds are included), of most interest are the factor analytic findings on subjects within this age range. Further, the subjects in the present study are all referred for evaluation due to some sort of learning and or behavior problem and so data from atypical populations will be reviewed.

The first study Nerviano reanalyzed was Wechsler's original standardization group data. For both ages 16 - 17 and 18 - 19 two factor solutions were obtained. The following factor loadings were observed (N = 200 per group). (Table 1)

The third factor (Freedom from Distractability) which is sometimes found in studies of the WAIS and WISC/WISC-R was not present in the standardization group. When it is found Digit Span Loads most likely on it. Arithmetic and Digit Symbol (Coding in the WISC/WISC-R) are sometimes found to have substantial loadings on this factor also.

Three of the studies dealing with a factor analysis of atypical samples will also be reviewed, since the present sample includes all three of these atypical samples to a greater extent than could be expected by chance.

Russell (1972) provided data for brain damaged (N = 77) plus normal subjects (N = 26). The factor loadings reveal the third

Table 1

Factor Loadings from the WAIS Standardization Group.

	Age			
	16 - 17		18 - 19	
Subtest	I	II	I	II
Information	89	40	88	69
Comprehension	84	40	82	53
Arithmetic	81	43	78	54
Similarities	82	47	86	64
Digit Span	70	38	74	42
Vocabulary	89	47	91	62
Digit Symbol	61	47	74	58
Picture Completion	71	88	66	83
Block Design	72	76	60	88
Picture Arrangement	68	70	60	80
Object Assembly	43	91	53	87

Note: Decimal points omitted.

factor to be present, but loading only on Digit Span. (Table 2)

The second factor (Perceptual Organization) is much more clearly delineated from the first factor (Verbal Comprehension) in this group than it was in the standardization group.

Spraque and Quay (1966) provide data on a retarded sample (N = 124). Once again the third factor is present. This time both Digit Span and Digit Symbol load highly on it. (Table 3)

In the case of this retarded sample it is the first factor that is not well delineated. The Comprehension, Arithmetic and Similarities subtests have almost equal loadings on more than one factor. Also notable is that the Similarities subtest has small loadings on all factors in this sample.

Finally, Paulsen and Lin (1972) provide us with data on a broadly based sample of inpatients and outpatients with a wide variety of psychiatric diagnoses (N = 290). A three factor solution was again obtained. (Table 4)

Again Digit Span provides the highest loading on the third factor. Arithmetic, as in the retarded sample has large loadings on all three factors.

Certain generalizations can be made from an analysis of these data. Four verbal subtests (Information, Comprehension, Vocabulary and Similarities) load highly on the first factor--with the exception of Similarities in the retarded sample. Of these, Vocabulary consistently loads the highest. Arithmetic tends to load substantially on more than one factor when a third factor is present. Digit Span loads on factor three in all but the normal standardization group. Four performance subtests load consistently higher on the second factor.

Table 2

WAIS Factor Loadings from a Reanalysis of Brain Damaged Subjects (Russell, 1972).

Subtest	I	II	III
Information	85	40	36
Comprehension	87	44	19
Arithmetic	81	60	33
Similarities	84	54	49
Digit Span	47	39	96
Vocabulary	91	39	37
Digit Symbol	38	84	39
Picture Completion	64	83	24
Block Design	50	85	38
Picture Arrangement	47	86	15
Object Assembly	37	87	24

Note: Decimal points omitted.

Table 3

WAIS Factor Loadings from a Reanalysis of Retarded Subjects (Spraque and Quay, 1966).

Subtest	I	II	III
Information	75	23	39
Comprehension	70	64	12
Arithmetic	57	50	54
Similarities	36	31	28
Digit Span	39	21	88
Vocabulary	80	19	15
Digit Symbol	19	56	77
Picture Completion	33	70	28
Block Design	28	79	33
Picture Arrangement	51	60	23
Object Assembly	16	82	23

Note: Decimal points omitted.

Table 4

WAIS Factor Loadings from a Reanalysis of Psychiatric Patients (Paulsen and Lin, 1972).

Subtest	I	II	III
Information	86	43	43
Comprehension	81	51	39
Arithmetic	64	56	65
Similarities	84	47	24
Digit Span	56	33	84
Vocabulary	89	36	50
Digit Symbol	51	79	24
Picture Completion	71	73	-06
Block Design	49	82	30
Picture Arrangement	65	76	10
Object Assembly	35	84	00

Note: Decimal points omitted.

Of this group, Object Assembly and Block Design (in that order) have the highest loadings followed by the Picture Completion and the Picture Arrangement subtests. The Digit Symbol subtest (a rote copying task) is the most inconsistent of all. It loads most highly on the first factor in the standardization group and on either factors two or three in atypical groups. This inconsistency, as we shall see, is found on studies of the WISC and WISC-R also.

Now that we have an idea of what to expect from a factor analysis of the WAIS, we will turn to studies of the WISC and WISC-R that have compared the factor structure of blacks and whites.

Factor analytic comparisons on the WISC and WISC-R. We will now review studies of the WISC and WISC-R which have compared the factor structure for blacks and whites. The purpose in doing this is to get an idea of the kinds of differences that have been found in order to develop hypotheses for the present study. Before beginning some differences in terminology will be noted. Digit Symbol on the WAIS is called Coding on the WISC and WISC-R. In addition, Mazes is a supplementary subtest of the WISC and WISC-R.

The problems mentioned by Nerviano when trying to summarize the factor analytic literature on the WAIS are present, but to a lesser degree, in studies of the WISC and WISC-R.

The first study performed was one by Semler and Iscoe (1966). The Picture Arrangement, Comprehension and Mazes subtests were not included as they had not been administered to this sample. These authors collected WISC data on 141 white and 134 black children ages five - nine who had been referred by their schools for evaluation. An exploratory maximum likelihood factor analysis revealed five

significant factors for blacks and three for whites. However, it was felt that this was due to difficulties in running the program and so the data were rerun (twice). The final time a three factor solution for both groups was performed followed by a varimax rotation. The factor loading obtained are shown in Table 5.

This study was presented because it was the first that directly compared the factor structure for blacks and whites. While the authors remark on the similarities between the factor loadings for the two groups, they note some glaring (and unusual) discrepancies. Arithmetic loads highest on the second factor for whites but on the third factor for blacks. Picture Completion loads highest on the first factor for both groups, which by itself is unusual. Finally, Object Assembly loads very highly on the third factor for whites but on the second for blacks. It must be assumed that the difficulties the authors admit to having in getting the program to run properly might be at least partly responsible for these unusual findings.

The next two studies to be reviewed make use of normal children and random samples. They have the additional feature of greater N's than most other studies reported.

Gutkin and Reynolds (1981) analyzed the data from the WISC-R standardization sample. 1868 whites and 305 blacks were included. Separate principal component factor analyses with varimax rotations were performed for both two and three factor solutions. Both solutions were tried to keep this research consistent with past factor analytic studies of the WISC-R. The resulting factor was assessed with coefficients of congruence. In addition, similarities in the strength of each factor across race were measured by comparing the portion of

Table 5

WISC Factor Loadings from Semler and Iscoe (1966).

	White			Black		
Subtest	I	II	III	I	II	III
Information	696	244	034	536	227	314
Arithmetic	426	624	141	194	033	978
Similarities	623	148	054	767	048	171
Vocabulary	733	263	227	710	105	066
Digit Span	257	463	180	385	199	154
Picture Completion	426	149	265	570	394	112
Block Design	213	646	308	199	385	233
Object Assembly	177	326	920	068	973	041
Coding	100	544	075	294	309	-058

Note: Decimal points omitted.

total variance accounted for by common factor variance and the percentage of common factor variance accounted for by each factor.

Gutkin and Reynolds conclude that the magnitude of variance associated with each subtest and the pattern of factor loadings were extremely similar for the two groups. Whether a two or three factor solution were employed every subtest loading highest on one factor in the white group also loads highest on that factor in the black group. When the number of eigenvalues greater than one are used to determine the number of significant factors the third factor is barely accepted in the white group, and rejected in the black group.

The final study to be reviewed is one by Reschly (1978). This study was part of a larger one funded by the Division of Special Education of the Arizona State Department of Education. While equal numbers of four groups (N = 260 each) were used, we will only report the data for blacks and whites.

These authors had two criteria for determining the number of significant factors in each group. First, the eigenvalue greater than one criterion was employed. Second, separate maximum likelihood factor analyses for two-, three- and four-factor solutions were performed followed by chi-square goodness of fit tests to compare the resulting factor matrix with the original correlation matrix.

The authors concluded that a three factor solution was best for the white group and a two factor solution for the black group. However, the eigenvalue for the third factor of the black group (also the other two groups) was close to one and so a three factor principal components analysis with varimax rotation was performed.

Table 6

WISC-R Factor Loadings and Eigenvalues for each factor from Gutkin and Reynolds (1981).

	White		Black		White			Black		
Subtest	I	II	I	II	I	II	III	I	II	III
Information	72	28	72	23	63	26	35	58	21	42
Similarities	67	33	66	30	63	32	26	62	28	28
Arithmetic	58	26	63	22	37	21	55	34	16	68
Vocabulary	81	24	81	24	77	33	31	80	21	33
Comprehension	62	27	59	34	64	26	18	58	33	23
Digit Span	43	17	49	27	18	10	60	23	23	57
Picture Completion	29	53	36	56	31	53	09	38	55	13
Picture Arrangement	33	45	30	47	32	44	14	28	47	14
Block Design	33	73	32	68	23	72	30	19	67	32
Object Assembly	18	67	19	73	19	66	09	19	72	12
Coding	31	23	30	22	16	20	36	17	20	29
Mazes	16	45	18	54	09	44	20	10	53	19
Eigenvalue	4.9	1.2	5.1	1.3	4.9	1.2	1.0	5.1	1.3	0.9

Note: Decimal points omitted.

Table 7

WISC-R Factor Loadings from Reschly (1979).

	White			Black		
Subtest	I	II	III	I	II	III
Information	63	32	26	66	40	18
Similarities	59	26	26	59	41	13
Arithmetic	43	26	45	61	34	27
Vocabulary	74	23	12	75	20	16
Comprehension	64	22	21	71	24	09
Digit Span	35	02	40	49	08	36
Picture Completion	20	49	09	25	52	21
Picture Arrangement	20	53	00	29	53	24
Block Design	17	60	22	20	33	58
Object Assembly	07	59	18	10	17	58
Coding	12	16	40	33	20	22
Mazes	18	42	10	23	44	30

Note: Decimal points omitted.

The authors conclude that, with some discrepancies, the resulting factor loadings were very similar. The fact that Object Assembly and Block Design, which usually are the subtests with the highest loadings on the second factor, load on the third factor for blacks is not explained. This may represent an artifact of overfactoring.

The findings of the principal component analyses of the WISC-R for blacks and whites suggests that the test does seem to be measuring the same two factors in both groups. When a third significant factor is found it just reaches significance in the black group. In addition, the subtest factor loadings are, for the most part, similar. Therefore, the argument that this test is biased because it is not measuring the same factors or constructs in both groups is not supported by these findings.

Now that the review of factor analytic findings on the WAIS and black-white comparisons on the WISC and WISC-R are complete, the specific statistical technique to be utilized in stage one of the current investigation will be discussed. While this procedure and the procedures of stage two will be described in some detail, as they are not common ones, the technical aspects of these procedures are not.

Data analysis. The data analysis for stage one will involve answering the question of whether the factorial pattern of WAIS scores is the same for both black and white groups. The statistical analysis technique to be used is derived from Joreskog's (1969, 1970) general linear structural relations methodology. While this methodology has not been used extensively to date, Reynolds (1982) notes that it has obvious advantages for studies of test bias, since any group

differences in parameters of the factor analysis model can be statistically tested for significance, and the method will therefore probably be used more extensively in the future.

For purposes of the present analysis, if one can assume that a factor analysis model holds in each population, then any parameter in the factor analysis model (factor loadings, factor covariances, and unique variances) for the two groups may be assigned an arbitrary value or constrained to be equal to some other parameter (Joreskog, 1971). For example, certain subtest factor loadings were set to zero. This procedure will allow one to set up a sequence of models such that each one is a special case of the preceding and will therefore offer the opportunity to test hypotheses concerning the factorial structure of the test for both groups.

For example, let us assume that there are only two factors present in both groups and we wish to test the hypothesis that the correlation of the factors is not significantly different for the groups. One can constrain the factor correlation to be equal in both groups and then statistically test whether this constraint is reasonable given the observed correlation found in each group. The test statistic's value will depend on the discrepancies between the observed correlations and the expected correlations (which would be somewhere between the two under the constraint that they are equal). This procedure of setting up constraints and testing them can be systematically applied to the other parameters in the factor analysis model.

The test statistic for evaluating these hypotheses is a large sample chi-square statistic (Joreskog and Sorbom, 1979). This test

is a test of the hypothesis that the measurements are structured in a manner defined by the model and the restrictions of fixed, free and constrained parameters as against the hypothesis that the measurements are not so structured, that is, that the variance-covariance matrix is unconstrained in both groups.

In the proposed study models will be tested against more general models by estimating each of them separately and comparing their chi-square goodness of fit values. The difference of chi-squares is asymptotically a chi-square with degrees of freedom equal to the corresponding difference in degrees of freedom.

The specific study that is being proposed is similar to the one performed by McGaw and Joreskog (1971). Test data derived from the Project Talent study was analyzed by this method. McGaw and Joreskog showed that a model with the same number of factors and a restricted pattern of factor loadings (tests loaded on certain factors only) held for groups differing in ability and socioeconomic status.

It should be emphasized (Joreskog, 1971) that significance levels are unknown when a sequence of tests like these are carried out and that even if a chi-square is large (large chi-squares indicate the observed data does not fit the model) there may still be reasons to consider the model. After all, the basic model with its assumptions of linearity and normality is only regarded as an approximation to reality. The true population covariance matrix will not in general be exactly of the form specified by the hypothesis, but rather there will be discrepancies between the true population covariance matrix and the formal model postulated. Whether to accept or reject a model cannot be decided on a purely statistical basis. This is largely a matter of

the experimenter's interpretations of the data, based on substantive theoretical and conceptual considerations.

Hypotheses to be tested in stage one.

1. The same number of factors adequately reproduce the correlations among the subtests in each group.
2. There is an invariant but unrestricted factor pattern (subtests will load equally on the same factors in both groups.

   In addition to these hypotheses which have been formally proposed in this study, other hypotheses may be tested to further explore the similarities and differences in group factor structures.
3. A restricted factor pattern holds for both groups. In other words the six verbal subtests will load on only one factor, the five performance subtests will load on only the other factor.
4. Not only is the factor pattern invariant over groups, but the unique variances are also equal.
5. The whole factor structure (factor loadings, factor covariances, and unique variances) is invariant over groups, with the same restricted factor pattern as before.

Based on the results of these analyses, one will be able to make some initial statements about the charges of bias that have been brought against the WAIS. For example, if it is found that the pattern of factor loadings on the subtests differs significantly for black and white students, then this can be viewed as supportive of William's hypothesis that the test is not measuring the same psychological dimensions or factors equally in both white and black groups.

Stage Two

Introduction. The area of item bias detection has come into prominence in only the last few years and so the discussion of this stage will begin with some introductory comments.

The second stage of the data analysis will involve an investigation of the individual test items to identify any differential patterns of responding. Differential patterns of responding refers to the case in which whites and blacks of a given ability level have different probabilities of answering an item correctly. Consistent with this, the formal definition of item bias is the following: An item is biased if two individuals of equal ability, but of different group membership, do not have the same probability of success of it. This definition closely parallels the definition of test bias mentioned earlier and will allow a statistical evaluation of test items.

The various techniques for item bias detection that have been developed are means of identifying items that are being responded to differentially within the context of other items, and are largely restricted to the identification of relative bias. Ability is usually (with the exception of the three parameter latent trait model) estimated by the total score on the test or subtest. The assumptions underlying this use of total test score are that the test is homogeneous (measuring one ability only) and the total score on the test is a good estimate of a person's true score on that ability.

Item bias detection methods are most useful when the great majority of items comprising the test or subtest are unbiased. When this is not the case a difficulty arises. In the extreme instance, if all of the items on a particular test are equally biased then the

estimate of ability will be similarly biased and there will be no discrepant items found (items that are being responded to differentially).

Realizing this, Shepard, Camilli and Averill (1981) attempted to evaluate test items against an external criterion of ability. The external criterion of ability chosen was a performance test--Raven's Progressive Matrices--that was free of verbal content (except for the brief instructions) and hence was assumed to be a less biased criterion. These authors felt that utilizing an external criterion of ability would avoid the potential confounding due to the fact that a biased item will bias the total score on the test containing that item. Unfortunately, they noted difficulties with their external criterion (pronounced ceiling effects) that mitigated their results. Nevertheless, these authors provided an idea that is well worth exploring. If an external criterion of ability that is assumed to be less biased could be found that correlates highly with the ability measured by the particular test or subtest, then evaluating items against this external criterion has definite advantages for an investigation of item bias. This is especially the case when one is trying to gether evidence to answer charges of bias that have been made against a test.

In the present study, ability will be estimated both by total subtest score and by a criterion that is external to the verbal subsets. This will enable one to compare the two procedures (Shepard, Camilli and Averill's intended purpose in including an external criterion) and also to present logically more defensible conclusions regarding item bias in the WAIS.

Item bias studies of the WISC and WISC-R. As has been noted earlier there are no item bias studies of the WAIS to date and therefore studies of the WISC and WISC-R that have compared blacks and whites on individual test items will be reviewed. Doing so should help one to develop hypotheses concerning a black-white comparison of the items on the WAIS.

Three of the studies to be reviewed make use of an ANOVA methodology in which items bias is determined by the finding of a significant item by group interaction. According to Cotter and Berk (1981) a primary concern of such studies is whether or not a significant item by group interaction is a reflection of bias or merely a reflection of initial ability differences between the subpopulations being studied. Hunter (1975) indicates that such evidence of bias is in actuality an artifact of the different ability levels of the two groups and hence different levels of item difficulty present. On easy items the black and white difficulty levels are similar, but as the items become more difficult the black difficulty level drops considerably faster than the white. Rather than representing an item by group interaction this finding may, as Hunter alludes, merely be an artifact of the different levels of ability of the groups. Another difficulty is that two of these studies used the original raw scores. These scores, which are mostly 0,1 or 0,1,2 are not interval in nature as the ANOVA model requires.

The first three studies all appeared in the year 1979. By the way in which they quote each other one can estimate which came first.

Miele (1979) analyzed data from a longitudinal study of 163 white and 111 black youngsters who had been administered the WISC in

kindergarten, first, third and fifth grades. There was some attrition over the five years so that at age eleven there were only 128 white and 97 black children. The average difference in group mean full scale IQ scores was 18.6 points, which, while higher than that found in the standardization groups, is typical of blacks and whites in the southeastern United States (the study was performed in Georgia).

Miele performed an ANOVA and found that the item by group interaction was significant beyond the .01 level for every subtest at every age. However, he reports that the amount of total variance explained by this interaction was no more than 5% and further the rank order of item difficulties in both groups was very high ($r = .96$). On further analysis, Miele found that the items which best discriminated between blacks and whites at any age level were also the items which best discriminated between age groups within race. From this Miele suggests that races might differ in mental maturity and differences between them are therefore not due to biased testing instruments.

Oakland and Feigenbaum (1979) compared groups of 180 white, 119 black and 137 Mexican-American children aged seven through fourteen who had been administered the WISC-R (among other tests). These authors measured test bias by internal consistency, indices of item difficulty, item-total correlations, concurrent and construct validity (factor analysis). They found that the groups differed in item difficulty only, and no consistent pattern to these items could be found that might be reflective of bias. Therefore, they conclude that the differences in difficulty level alone were probably a reflection of the differences in ability.

Sandoval (1979) analyzed WISC-R data from equal numbers (350 each) of white, black and Mexican-American children five through eleven years of age. These children were part of the SOMPA standardization group (Mercer and Lewis, 1978). Two indices of ranked item difficulty, ANOVA of items x race (SES nested within race) and MANOVA of subtests with Race x SES were performed. Sandoval's analysis differed from the others previously mentioned in that subtests were analyzed individually.

Sandoval considered the MANOVA to be a powerful way of determining which items contribute to the observed differences in ethnic group performance. Following the MANOVA performed on the difference in item means across ethnic and SES groups for each subtest, Sandoval analyzed the eigenvalue or amount of explained variance, the canonical correlation, and the significance of the function for each effect. According to Sandoval, those items most discriminating between groups have correlations with the canonical variate that are greater than .40. A total of 45% of the 176 items studied met this criterion. However, only two of the subtests as a whole--Vocabulary and Information--showed significant black-white differences.

When content analyzed, Sandoval, as had Oakland and Feigenbaum, could find no discernible pattern to these items, and concludes that explanations other than specific item content must be explored for explaining black-white differential performance. Following Sattler (1979) Sandoval suggests that minority groups may be deficient in motivation, test practice and reading skills, and these, not test bias, may be the reason behind lower minority scores. In the opinion of this writer only motivation, of these, represents a plausible explanation for group differences on the WIAS since items are read

to subjects and the subtests are not of the variety that a person is likely to have had much practice with.

The final study to be reviewed represents a substantial advance over the previous ones reported. Cotter and Berk (1981) attempted to add precision to the ANOVA model in their study. They collected WISC-R data on 112 black, 126 white and 117 Hispanic learning disabled youngsters between 10 and 10.33 years of age in a large metropolitan county in Florida.

To control for differences in ability the authors matched groups on subtest raw score (they were also matched on sex). Following this, item means were then transformed first to a z scale and then to a delta scale. A delta scale is a normalized scale with a mean of 13 and a standard deviation of 4. This was done because delta transformed item means approximate an interval scale of measurement (Jensen, 1980) and therefore more powerful statistical analyses can be performed than on raw item means which are not on an interval scale. Following this an item by group repeated measures ANOVA was computed on delta transformed item means for each of nine subtests and two group comparisons to determine interaction effects. Significant interactions were then followed by post hoc comparisons with the Newman-Keuls test.

The authors found that five subtests (Information, Similarities, Vocabulary, Comprehension and Picture Completion) contained items that were biased against either whites, blacks or hispanics. A total of 18 items or 22% of the items in these subtests were found biased by their criterion. When only the black and white groups were compared 20% of the items in four subtests (Information, Similarities, Comprehension and Picture Completion) were found biased. However, the

direction of bias tended to balance out. 11% of these items were biased against blacks and 9% were biased against whites. Therefore, no one group was consistently at a disadvantage.

While this study does represent an improvement over the earlier studies mentioned, the findings can be questioned because of the use of subtest total score to match groups on as a means of controlling for ability. The subtest may contain biased items and if more of these items are biased against blacks then the subtest total score would tend to underestimate the actual ability of blacks on the ability being measured by the subtest. Whites and blacks matched on total test score would than represent higher ability blacks compared with lower ability whites. If these two groups have the same probability of answering an item correctly then the item is really biased against blacks because blacks should do better if the item were not baised. It was because of this argument that Shepard, Camilli and Averill (1981) suggested the use of an external criterion of ability as a means of avoiding this potential confounding.

Item bias detection procedures. Various item bias detection procedures have been presented in recent years. Of the many approaches there are four that have received the most attention: item discriminations, transformed item difficulties, chi-square procedures and latent trait models. The usefulness of these procedures, among others, for purposes of detecting item bias has been studied (Ironson and Subkoviak, 1979; Rudner, Getson and Knight, 1980; Shepard, Camilli and Averill, 1981). These investigators all agree that the most theoretically sound technique, which it is felt also happens to be

best for identifying biased items, is the three parameter latent trait model (ICC-3). This is followed by the chi-square procedures, which are recommended for use when latent trait models are not appropriate. While it is one of the chi-square procedures that will be used in the present investigation, both of these procedures will be discussed since the chi-square procedures are attempts to approximate the more sophisticated latent trait models.

Latent trait models. The use of latent trait theory or item characteristic curves (ICCs) in the evaluation of test items has become prevalent in the past several years (Lord, 1980; Wright, 1979). More recently latent trait theory has been applied to investigate the issue of item bias (Ironson and Subkoviak, 1979; Rudner, Getson and Knight, 1980; Shepard, Camilli and Averill, 1981).

Using this method, observable scores on test items are viewed as functions of an unobservable trait or ability which determines these scores. ICCs describe the relationship between an examinees ability and the probability of a correct response to a test item. The item and ability parameters are estimated by maximum likelihood methods and require that two assumptions be met: unidimensionality and local independence. The former is a requirement that the test measure a single underlying ability only. The latter refers to the assumption that an item on a test does not determine his or her responses to other items on the test.

With this technique an item is considered to be unbiased if examinees of the same ability level, but of different group membership, have equal probabilities of providing a correct response to the item. The benefit of using latent trait models is that they are said to

offer sample invariant estimates of latent trait parameters. While Wright (1979) has shown this to be true for the Rasch model which deals with one parameter only (item difficulty), Shepard, Camilli and Averill (1981) note that this has not yet been so shown for the additional item discrimination and item guessing parameters of the three parameter model. These authors do conclude, however, that the three parameter model is not as effected by differing group ability levels as are the other item bias detection procedures.

Three parameter cumulative logistic model (ICC-3). The ICC-3 model due to its theoretical sophistication and especially for its capability of handling ability as a continuous variable is thought to be superior to other procedures for detecting item bias.

The model, as described by Birnbaum (1968) is a logistic function which relates item characteristics and ability to the conditional probability of passing an item. The more ability an individual has the greater the probability of their responding correctly to an item. Three parameters are estimated in this model: An item discrimination parameter (Ag); an item difficulty parameter (Bg), and an item guessing parameter (Cg). Lord (1980) has developed a significance test for estimating the difference between group item parameters.

While this model has obvious theoretical and practical advantages for detecting item bias it also has serious drawbacks that make it impractical to use in most cases. The most serious drawback is the need for large sample sizes (at least 1,000 per group) to achieve stable parameters. In addition, it requires the test to have a minimum of forty items (Scheuneman, 1970). Finally, it is sensitive to violations of the assumptions underlying its use.

For the following reasons the model is, unfortunately, not thought appropriate for the present study: First, sample size only averages 300 per group. Second, subtests must be evaluated separately and six out of nine subtests have less than fifteen items, and finally, many of the items are not scored dichotomously. The three parameter model, as currently used, requires items to be scored dichotomously. In addition to these drawbacks, all past item analyses of the Wechsler scales violate the assumption of local independence and therefore must be viewed as exploratory.

Test administration procedures require one to stop administering many of the subtests after a set number of incorrect responses (usually four or five). The rationale behind this is that items are supposedly arranged in order of difficulty, and if a person fails four easier items he or she is not expected to pass the harder items. Therefore, the scores of latter items depend somewhat on scores of earlier items. While this is definitely a problem to be considered, it is fortunately not a serious violation of the assumption of local independence.

Chi-Square Procedures

Scheuneman's chi-square method. Scheuneman (1975) has provided us with a methodology utilizing a chi-square procedure for detecting biased items. Her methodology was designed to be a rough approximation of the more sophisticated latent trait models.

The rationale for her procedure is based on a definition of item bias that is similar to the one proposed by latent trait model theorists: "An item is unbiased if, for all individuals having the same score on a homogeneous subtest containing the item, the proportion of individuals getting the item correct is the same for each population group being considered."

To achieve the goal of equating subjects on ability total score on the test is divided into discrete intervals, usually three to five. Care must be taken in specifying the intervals for each item so that there are sufficient cases in each interval and the probability of answering correctly is less than one and greater than zero even in the extreme ability intervals. Further, when ability intervals are made too broad to obtain sufficient cases, the group abilities may differ and spurious results may occur.

An expected set of proportions is calculated assuming ethnic group and the probability of a correct response to be uncorrelated in each ability group (only correct responses are analyzed with this procedure). Then the observed proportions are compared with the expected proportions using a modified chi-square statistic. Large deviations from expected frequencies, summed across intervals will result in large chi-square values signifying a biased item. In studies with both simulated data (Rudner, Getson and Knight, 1980)

and with actual data (Ironson and Subkoviak, 1979) this procedure correlated better with the ICC-3 results than did any other procedure.

Full chi-square method. Camilli (cited in Shepard, Camilli and Averill, 1981) explained a logical flaw in the Scheuneman chi-square method. He demonstrated that chi-square statistics computed from proportions correct (p) across ability strata for different groups will lead to different results than if the chi-square were calculated using proportions of incorrect responses (q). Since the proportions of correct and incorrect responses should reflect the same information, this finding is anomalous and troublesome.

Camilli recommended using the more conventional full chi-square method (fitting expected values for both the correct and incorrect responses) over ability intervals and then summing across ability intervals. As with the Scheuneman procedure, large deviations from expected frequencies within ability strata will lead to large bias indices.

As a further refinement of the technique, Shepard, Camilli and Averill (1981) obtained both signed and unsigned chi-square values. Signed indices are obtained by inspecting the direction of bias in each interval and ascribing a sign before summing the squared discrepancies (Ironson and Subkoviak, 1979). Signed and unsigned indices will result in the same conclusions about the magnitude of item bias if one group is consistently disadvantaged. However, when discrepancies from expectation are large but compensating as is the case when item characteristic curves cross, then the signed index may not reflect bias found by the unsigned chi-square.

Shepard, Camilli and Averill (1981), compared sixteen different bias indices and concluded that the signed chi-square approach is preferred when the ICC-3 model cannot be used. The correlation of items identified as being biased by these two approaches averaged .70 in their study.

In the opinion of this writer the signed chi-square approach offers a more restrictive definition of item bias by minimizing chance fluctuations from expectations. However it also has a drawback. The drawback is that there may indeed be interactions between groups and ability levels. Instead of identifying such interactions they would be minimized with this approach and therefore important information would be lost.

Log linear models and the likelihood ratio chi-square. Several investigators (Camilli, 1979; Marascuilo and Slaughter, 1981; Alderman and Holland, 1981) have suggested the use of a third chi-square methodology--log linear models--as an efficient way to analyze contingency tables for bias. The responses to each item can be conceptualized in a three way table: Ability level x racial group x response category (correct-incorrect). The fit of various models to the observed data can then be tested. If a model with a constant and ability level can adequately fit the data, there is no bias. If a term including racial group is necessary in addition, then there is uniform bias (for all ability categories, the difference between response ratios of blacks and whites is constant). If that model does not fit the data, a parameter for the interaction between ability group and racial group is needed, and the item is nonuniformly biased.

Since the technique is not as well known as are other chi-square procedures we will now describe it more fully.

Traditional chi-square procedures based on the Pearsonian statistic allow us to test one hypothesis: The cells in a contingency table are independent. Goodman (1978), a major proponent of log linear models, offers another way of analyzing qualitative data that has definite advantages for a study such as the one being proposed. In essence the procedure involves computing the natural log odds of being in a category and comparing this to the observed odds. Expected versus observed odds are then tested by means of a likelihood ratio chi-square statistic.

As an example let us consider the simple model:

		B: 1	B: 2
A	1	F_{11}	F_{12}
	2	F_{21}	F_{22}

F11/F12 gives us the odds that B equals 1 given A equals 1.

F21/F22 gives us the odds that B equals 1 given A equals 2 etc.

Log linear analysis allows one to test many more hypotheses than the traditional Pearsonian chi-square approach. In the above example the following models can be tested:

HO: cells are independent.

H1: A main effect of A is present.

H2: A main effect of B is present.

H3: Both A and B main effects are present but independent.

H4: Both A and B main effects and an A x B interaction effect are present.

This last model cannot be tested directly because there are no degrees of freedom left to evaluate the chi-square statistic.

Any given model can be investigated by comparing it to a less restrictive model. The difference in their likelihood ratio chi-squares is distributed as chi-square with degrees of freedom equal to the corresponding difference in degrees of freedom.

While log linear models are more complex than the traditional chi-square analysis a number of computer packages have incorporated it into their format (BMDP-3F, MULTIQUAL, SAS-FUNCAT).

Log linear models have been utilized in a recent study of the Test of English as a Foreign Language (Alderman and Holland, 1981). These authors conclude that the application of this technique is straightforward and efficient. In addition it leads itself readily to the examination of multiple groups and multiple dimensions, and finally, it has the decided advantage of allowing one to estimate higher order or interaction effects. Alderman and Holland therefore recommend the use of log linear models for detecting item bias.

Advantages and disadvantages of chi-square procedures. The advantages and disadvantages of the chi-square procedures for detecting item bias will now be discussed. This section owes much to Ironson (1982). First those advantages and disadvantages that the methods share in common will be mentioned and then the advantages and disadvantages specific to each procedure will be evaluated.

Advantages

1. The procedures are intuitively understandable to the practitioner, as they attempt to answer the basic question: Do different groups of the same ability have the same probability of getting the item right.
2. Smaller sample sizes (N of 100 - 200 per group) are required than with the latent trait models which these procedures attempt to approximate.
3. The procedures are associated with significance tests that permit dichotomous decisions biased/unbiased to be made.

Disadvantages

1. Problems in using the total test score as a measure of ability. The procedures assume the total test score is a valid measure of ability. While total test score is a reasonable choice, it does have several weaknesses. First, the total test may contain biased items. Suppose, for example, that there is more bias against blacks than against whites. The total test score would then underestimate black ability. Whites and blacks matched on total test score would then represent higher ability blacks compared with lower ability whites. If these two groups have the same probability of answering an item correctly, the item is really biased against blacks because blacks should do better if the item were not biased. In addition, constant bias (all items are equally biased) may be absorbed into the scales; thus by using total test score as a measure of ability, constant bias will be overlooked. A possible remedy for this is to utilize an unbiased or at least a relatively less biased external

criterion of the ability measured by the test (subtest) to establish the intervals with. Second, sometimes groups in the same total test score interval may not have the same mean test score. This will occur if the distributions of total test scores within an interval are different, especially at the highest levels of ability. Therefore, groups may not really be matched on ability. Due to differences in group means within "matched" intervals, regression artifacts may still cause the appearance of bias. By checking the distributions of scores in intervals of items identified as biased one can eliminate such unwarranted instances.

Finally, Ironson (1982) notes that the total test score is not perfectly reliable (especially with tests with few items). One potential way of overcoming this is to equate groups on a valid, more reliable measure of the ability measured by the test.

2. Arbitrariness in setting cutoffs for the ability intervals. Within the constraints set for the procedures by the required expected frequencies, there are a variety of possible cutoffs. The magnitude of the chi-square could change as the cutoffs change.
3. Failure to utilize all available information. None of the procedures focus on the continuous, quantitative nature of the underlying variable (Marascuilo and Slaughter, 1981). Instead, the underlying variable (ability) is treated essentially in categorical fashion.
4. Sensitivity to differential sample size of groups. Baker (1981) noted that the Scheuneman chi-square procedure is confounded by

unequal sample sizes for the two groups. He demonstrated that "two sets of identical pseudo item characteristic curves could lead to different chi-square values and possibly different conclusions concerning item bias if they were based upon different group size ratios." Comparing equal sample sizes(N = 800), he obtained a Scheuneman chi-square of 11.814 (3df, $p < .05$); with a ten to one ratio of sample size (N = 440) it dropped to 1.80 (3df, $p < .05$). To avoid this, Shepard, Camilli and Averill (1981) recommend randomly removing cases if sample sizes are too unequal. Before resorting to this, however, they recommended that attempts should be made to collect data on the smaller group.

Scheuneman's chi-square correct

Advantages.

1. Smaller sample sizes are required than by the chi-square full procedure. This is because it is not necessary to have expected frequencies of five for the incorrect responses. Due to this, more items can be evaluated with the chi-square correct procedure (easy items) than with the chi-square full procedure.
2. It is less sensitive to contributions from different cells. Scheuneman's chi-square correct is relatively insensitive to certain contributions from the top ability group. Consider, for example, an easy item. For high ability people, answering it incorrectly may be largely random. Since the expected frequency appears in the denominator of the contribution to the chi-square, the contribution is likely to be large. Therefore, random variation would be receiving a lot of weight if the incorrect responses were included. Added to this is the concern that the

top ability groups are not really matched on total test score. The whites in that category typically have a wider variance in total test score, whereas blacks tend to be concentrated toward the lower cutoff. One could argue that the reverse problem is present for using correct responses in the bottom ability category, particularly for difficult items. However, the situation is not symmetrical because the expected frequencies are not as low, due to guessing, and the difference in total test score (mean) in the low ability group is generally much smaller, particularly since the sample size tends to be smaller for the low ability group.

Disadvantages

1. Both Scheuneman and other investigators have noted that the chi-square correct is not distributed as a chi-square statistic. At present, the sampling distribution of this statistic is not known.

Chi-square full

Advantages

1. The chi-square full procedure is distributed as a chi-square statistic.
2. Computer programs are ready available.

Disadvantages

1. A larger sample size is required than with the chi-square correct procedure.
2. It is more difficult to evaluate easy items.
3. It is more sensitive to random variation in high ability groups (see advantages of Scheuneman's chi-square correct procedure).

Log linear models

Advantages

1. The test statistics are distributed as chi-square.
2. A greater number of models (main effects and interactions) can be tested simultaneously.

Disadvantages

1. It is more difficult to interpret than traditional chi-square approaches.
2. Not as many computer packages include log linear models as include the full chi-square procedure.

Hypotheses to be tested in stage two.

1. Black and white students of equal ability will have the same probability of answering each item correctly (no biased items).
2. If items are found for which black and white students of equal ability do not have the same probability of answering correctly, then the direction of biased items will tend to balance out so that no one group is disproportionately at a disadvantage.

Footnotes

[1]According to Jensen (1980), internal criteria of bias are considered only in terms of how particular statistical properties of a test differ significantly in any two populations. Judgemental item review procedures associated with the determination of content validity and, in this case, content bias are ignored. While these two procedures do not always result in the identification of the same items as biased, investigators such as Scheuneman (1979) note that statistically showing two groups to be responding differentially to an item is not nearly as important as an understanding of why this is the case. This understanding can only be had by a judgemental review of the item content.

[2]The Wechsler Preschool and Primary Scale of Intelligence (WPPSI) will not be included, as the research literature on this instrument, is not sufficient for the present purposes.

Method

Subjects

The subjects for the present study were high school students from Region Four schools in Brooklyn, N.Y. Since Region Four covers approximately half of Brooklyn a wide range of socioeconomic areas were included.

The subjects had been referred to school psychologists for an evaluation of learning problems, behavior problems, or both. As part of this evaluation the WAIS was administered. All evaluations took place between January, 1980 and June, 1982. While most of the students were being evaluated for the first time, approximately 25 percent were students in special education classes who were being reevaluated. While no information concerning country of birth is available for the total sample, 28 percent (21 of 79) of the blacks and 13 percent (6 of 47) of the whites tested by the author were born outside the United States.

Data were obtained on 256 white (180 male and 76 female) and 358 black (259 male and 99 female) subjects. The subjects, for the most part, range in age from 16-20. However, a few 15 year olds were included. This was often because the WISC or WISC-R had been administered within the past year, and the school psychologist administered the WAIS to avoid a practice effect. Although standardized instructions do not call for telling subjects that they are being timed and that the faster they correctly complete the items the higher the score they will receive, the school psychologists in the present sample routinely do this.

Instrument

The Wechsler Adult Intelligence Scale (WAIS) is an individually administered intelligence test that is utilized with individuals at least 16 years of age (Wechsler, 1955). The test consists of six verbal and five performance subtests. The median number of items per subtest is 13.5, with eight of the subtests having less than 15 items. The individual subtests are described in Appendix A at the end of this paper.

Approximately 50% of the items are scored dichotomously; the rest are multipoint items. Item scoring varies not only between subtests, but within subtests as well. As an illustration, on the Block Design subtest it is possible to get one of three possible scores on the first two items, one of two possible scores on the next four items, and one of four scores on the last four items.

Data

WAIS test protocols were collected from 12 school psychologists. The protocol scoring, as well as age, sex and race were copied onto a coding sheet and the protocols were returned to the school psychologists.

It was noted that some school psychologists would leave out subtests in order to save time. No protocols were accepted that had more than one subtest left out. About twenty seven percent of the protocols accepted contained one missing subtest. Only three subtests (Vocabulary (V), Comprehension (C) and Object Assembly (OA) were ever missing. The breakdown on missing subtests is shown in Table 8.

Procedure

The preliminary data analyses were the computation of descriptive

Table 8

Percentage of Subjects with Missing Subtests by Race and Sex.

	White			Black		
Test	Male	Female	Total	Male	Female	Total
V	18.9	19.7	19.1	17.4	14.1	16.5
C	3.9	4.0	3.9	3.5	3.0	3.4
OA	6.1	5.3	5.9	6.2	4.0	5.6
Total	28.9	28.9	28.9	27.0	21.2	25.4

statistics and the production of graphs to see if the data were normally distributed, and if not, how comparable to data from normal populations it was.

Following this the investigation proceeded with an analysis of the factor structure of the test scores. First, an exploratory principal components analysis with a varimax rotation was performed in each group, as was an exploratory maximum likelihood factor analysis. The information gained helped to build confirmatory factor analysis models, which were then tested for their fit to the observed correlation matrices. Since no clear pattern of factors emerged, a series of confirmatory factor analysis models were then tested in which loadings of variables (subtests) were set to zero if their estimate was not more than two standard errors from zero. Once acceptable models had been found for both groups, simultaneous confirmatory factor analyses were performed on the subtest covariance matrices. The LISREL IV computer program (Joreskog and Sorbom, 1978) was used for these analyses.

When this had been completed, the individual test items were investigated for differential patterns of responding by blacks and whites. Because their formats are not appropriate, the Digit Span and Digit Symbol subtests were not included in this analysis. The Digit Symbol subtest consists of one item only, while the Digit Span subtest consists of two parts that are not homogeneous.

A series of log-linear analyses were performed to determine whether groups have the same probability of obtaining both correct and incorrect responses to the items of the nine subtests. Since one alternative explanation for the findings may be that the groups differ

in the abilities measured by the particular subtest, ability level was used as a control variable in the analyses. This technique has some similarity to the one recommended by Scheuneman (1975). However, Scheuneman's and related procedures usually use the total score on a test as a criterion of ability when evaluating the items making up the test. This procedure has a number of liabilities which have already been mentioned, and therefore may have limited usefullness in a study designed to answer charges of bias brought against a test.

Shepard, Camilli and Averill (1981) recognized this difficulty and suggested the use of an external criterion as a means of establishing ability intervals. Ironson (1982) also recommends utilizing an external criterion if one that is unbiased, or is relatively less biased than the total subtest score can be found.

The external criteria used in the present study were Performance IQ and Verbal IQ. Performance IQ is independent of the scores on the verbal items, and so was used as a criterion of ability when evaluating the verbal subtests. Verbal IQ is independent of the scores on the performance items, and so was used as a criterion of ability when evaluating the performance subtests. Although Verbal IQ and Performance IQ are scores that are each external to parts of the WAIS, they are both internal to the test itself. They will therefore be termed <u>independent internal criteria</u> of ability rather than external criteria.

These scores were chosen as estimates of ability since the performance score (corrected for attenuation) correlates .73 with the verbal subtests in the standardization group, and the verbal score (corrected for attenuation) correlates .72 with the performance

subtests. Therefore these scores are good estimates of the abilities being measured by the individual subtests, and their use avoids the potential confounding inherent in using the total subtest score. This is deemed especially important in the case of the verbal subtests, since it is these that have been the subject of much of the criticisms regarding bias in the WAIS. Verbal subtests have been critized both for the content of the items favoring whites and also because it is felt that blacks have had less practice with the skills necessary to correctly answer the items. Performance subtests, on the other hand, which have little verbal content, have been criticized only for the latter reason and even these arguements are not as strongly made. Thus, Performance IQ is not only an independent criterion of ability for evaluating the verbal items, but is thought to offer a more defensible criterion than total subtest score when attempting to evaluate item bias. The same logical argument does not hold for Verbal IQ. Verbal IQ was used primarily as an independent internal criterion of ability for evaluating the performance subtest items.

For each item five ability intervals were set up on the basis of both an internal (total subtest score) and independent internal (Performance IQ or Verbal IQ) criterion of ability. All analyses were done with both internal and independent internal criterion so that the results of these two procedures for establishing ability intervals could be compared. This comparison had been one of Shepard, Camilli and Averill's (1981) intended purposes.

The ability intervals were chosen in the following way: The lowest score in the first (lowest) ability interval was the lowest score obtained on the criterion by both groups. The highest score

in the fifty (highest) ability interval was the highest score obtained on the criterion by both groups. While this procedure resulted in the dropping of certain extreme scores, and hence reducing the data base, it also helped to minimize the potential difficulty pointed out by Ironson (1982). Ironson notes that the mean scores, especially in the extreme ability intervals, are often different when blacks and whites are compared and therefore the groups are not really equated on ability. Further, for each criterion of ability, the width of each interval was the same for each item within a subtest.

If it was found that a particular cell frequency was less than five, that cell was collapsed with an adjacent cell. When cells required collapsing it was always because of small frequencies in the extreme intervals. A number of items were evaluated with more than one correct score category. If the observed frequencies in one of these correct score categories was too small for the analysis, that score category was collapsed into the adjacent correct score category. In no instance was an item evaluated if less than three intervals could be produced by either the internal or independent internal criterion of ability. This was only the case in items of extreme difficulty.

Once the biased items had been identified the task became one of trying to determine the direction of bias. This was determined by computing signed chi-squares (Ironson and Subkoviak, 1979) and comparing the value of the signed chi-square to the chi-square distribution. The sign is determined by examining the observed and expected frequencies and noting which group achieved more correct or less incorrect responses than could be expected by chance. A plus

sign denoted the fact that the white group obtained more correct or less incorrect responses than could be expected by chance. A minus sign denoted the fact that the white group obtained less correct or more incorrect responses than could be expected by chance. The contribution to the signed chi-square was summed over score categories when evaluating the direction of bias for a given ability interval, and summed over ability intervals when evaluating the direction of bias for the whole item.

The signed chi-square is not distributed as a regular chi-square unless all of the contributions to the signed chi-square have the same sign. When the signs differ then part of the contribution to chi-square is subtracted rather than added. Signed chi-squares therefore actually represent a more restrictive definition of item bias than regular chi-squares (Shepard, Camilli and Averill, 1981).

If the signs are not all the same, then the signed chi-square is not the same as a regular chi-square and probability levels cannot be determined. In this case, if the signed chi-square surpasses the value of the regular chi-square (with the same degrees of freedom) needed for significance at the .05 level, then it was said that the signed chi-square favors one group or the other. This is, of course, a conservative criterion.

A number of the items analyzed are scored other than correct or incorrect only. For example, on the Comprehension, Similarities, and Vocabulary subtests correct answers may receive either one or two points depending on the quality of the answer. In the present study two point responses were considered high quality responses, zero and one point responses low quality. Further, certain items of the Picture

Arrangement and Block Design subtests, and all of the items of the Object Assembly subtest, offer bonus points for correct answers that are performed within certain time limits. (Actually more than one level of bonus score was possible, but it proved necessary to collapse all bonus score levels into one correct plus bonus category.) Correct plus bonus point responses are here considered high quality responses. Incorrect and correct without bonus points are considered low quality responses.

The direction of bias for items with each of the above scoring characteristics was studied by performing signed chi-squares for each of two comparisons, since both of these comparisons represent instances of bias. The first comparison involved studying whether or not the two groups, when equated on ability, had the same probability of obtaining correct and incorrect responses. The second comparison involved an investigation of whether or not the two groups, when equated on ability, had the same probability of obtaining high quality and low quality responses.

Once this had been completed, an attempt was made to note any patterns of biased items. While other researchers (Oakland and Feigenbaum, 1979; Sandoval, 1979) had not had any success in analyzing the content of biased items for noticeable patterns, many potential patterns have not yet been explored. One of these patterns, which may involve motivation to achieve higher scores, was investigated in the present study. Specifically, personal experience seems to suggest that blacks and whites of relatively equal ability may not have the same probability of achieving correct plus bonus point responses on the performance subtest items. Since bonus point responses require

rapid performance they indicate not only the ability to correctly solve a problem, but the motivation to do so rapidly. In this regard, Matsui, Okada, and Kakuyama (1982) found a significant positive correlation between socres on the Manifest Needs Questionnaire and the number of items atpempted on a speeded perceptual task. The authors conclude that those highest in the need to achieve are the most motivated to work rapidly. Sattler (1979) has hypothesized that blacks may be more poorly motivated for test taking, and that this may be part of the reason for their receiving lower scores on intelligence tests. Therefore, the hypothesis that blacks and whites of roughly equal ability do not have the same probability of achieving correct plus bonus point responses was investigated. For each biased item containing bonus point responses the following procedure was carried out: First, correct plus bonus point responses were collapsed into the correct response only category (comparison one). The items previously detected as biased were then reanlayzed by log-linear analysis. If the discrepancy in bonus points is the major contributor to the statistical bias, then when this category is no longer included, fewer items should be found to be biased. Second, the direction of bias was noted in the high quality comparison of items containing bonus points. If the hypothesis that whites are more motivated than blacks for test taking is correct, then the whites should be favored in more ability intervals than blacks.

Results

Before discussing possible bias in the WAIS, various descriptive statistics will be presented to show how the two groups compare. The more alike the groups in terms of ability, age, and so on, the more confidence can be placed in the findings.

Table 9 shows that the two groups had similar age distributions.

Table 10 reveals some interesting results. First, the white mean Full Scale IQ score is approximately 10 points below that usually found in normal groups. The black mean Full Scale IQ Score, however, is similar to that typically found. While the mean for the white group is significantly higher than the mean for the black group ($t(612) = 3.75$, $p < .01$), the absolute difference (3.5 points) is much smaller than that often found in more representative samples. Further, while males and females in normal groups have similar full scale scores, in this instance male full scale scores are about 5 points higher than females, regardless of race. This finding also represents a significant difference in means ($t(612) = 5.05$, $p < .01$). The implications of these findings, particularly in relation to selection, will be discussed later.

Figure 1 depicts the distribution of Full Scale IQ Scores for the black group. It can be seen that the distribution of scores resembles a normal distribution.

Figure 2 represents the distribution of Full Scale IQ Scores for the white group. This distribution only much more roughly approximates a normal distribution. This also will be discussed later.

Besides maturation level and overall ability, another area that may influence the findings is in the subtest and summary score

Table 9

Mean and Standard Deviation of Age in Months for Males and Females Within Race and as a Combined Total.

	White			Black		
	Male	Female	Total	Male	Female	Total
$\overline{X}$	293.65	293.93	203.73	203.46	205.61	204.05
SD	10.52	11.66	10.85	11.41	11.84	11.55

Table 10

Means and Standard Deviations of Verbal, Performance and Full Scale IQ Scores for Sex Within Race, and as Combined Racial Groups.

		White			Black		
		Verbal	Performance	Full Scale	Verbal	Performance	Full Scale
M	$\bar{X}$	89.82	83.18	90.66	86.53	89.54	87.04
	SD	11.70	13.63	12.25	10.63	11.29	10.55
F	$\bar{X}$	86.24	86.15	85.46	83.34	82.36	81.81
	SD	11.16	13.93	12.46	8.87	12.01	9.99
C	$\bar{X}$	88.76	91.09	89.12	85.65	87.56	85.60
	SD	11.64	14.06	12.52	10.26	11.92	10.65

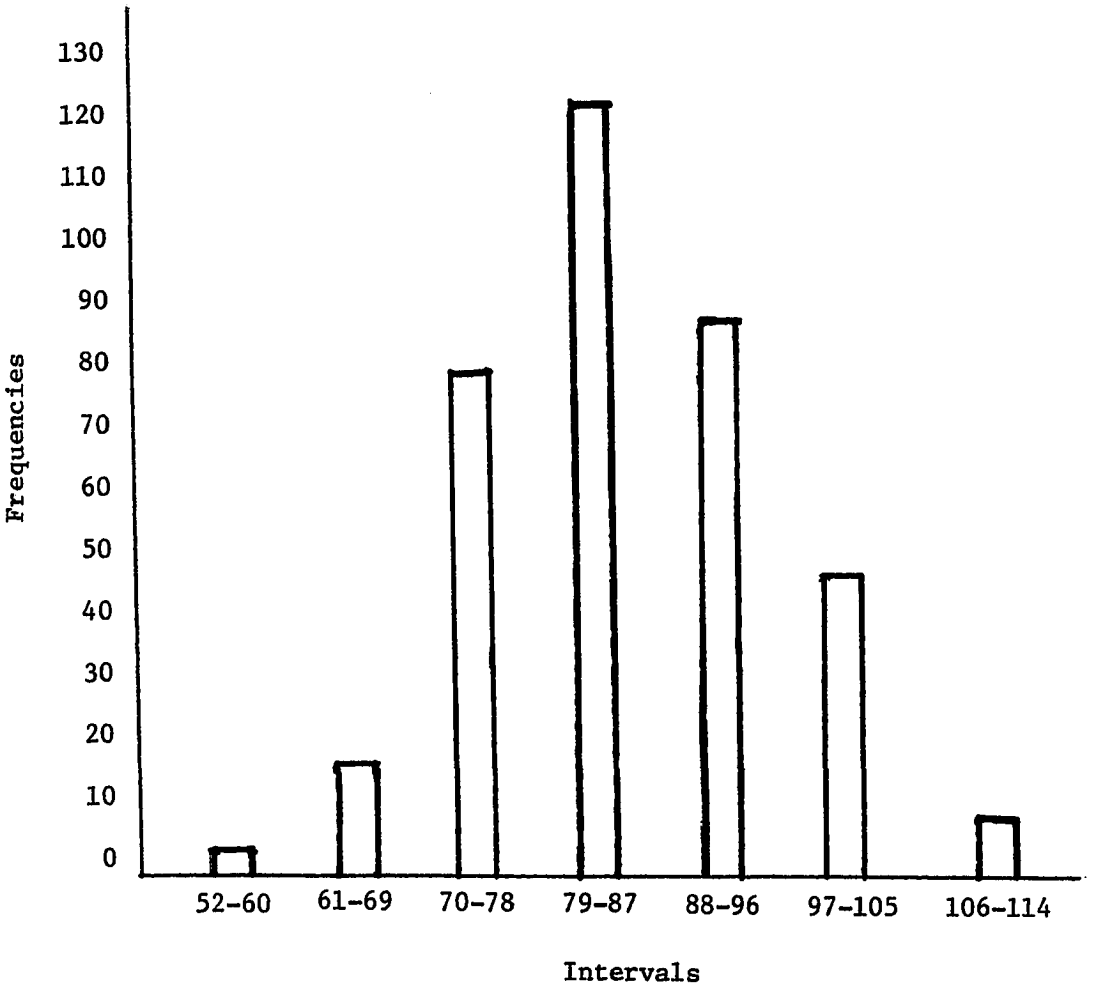

Figure 1. Frequencies of Full Scale IQ's for blacks.

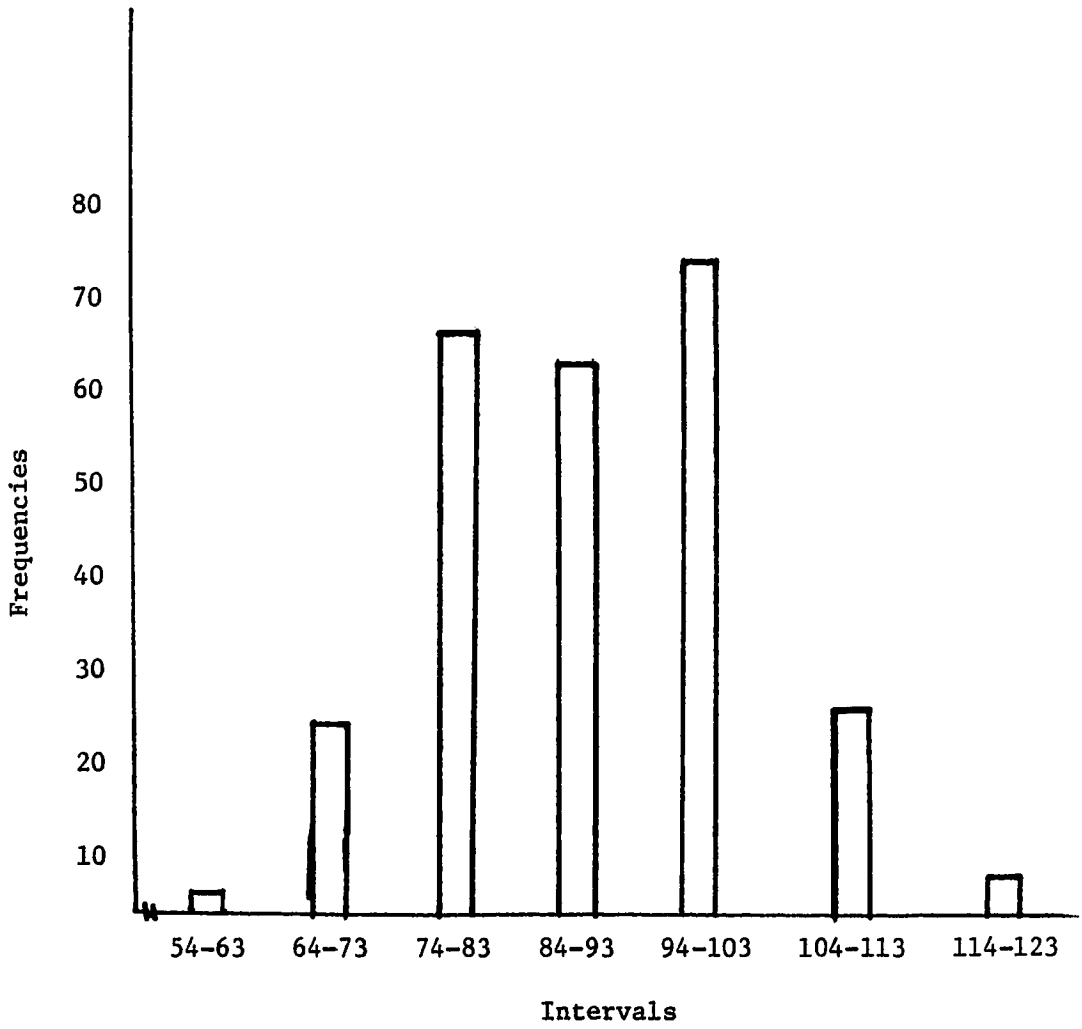

Figure 2: Frequencies of Full Scale IQ's for whites.

reliabilities. Coefficient Alpha's were computed for each subtest and summary score to estimate reliability. Item raw scores were used in the computations. Only complete protocols from 267 blacks and 182 whites were used for this purpose, and when computing correlation matrices for the factor analyses.

An analysis of Table 11 suggests that the reliabilities are comparable but somewhat higher for the white group.

The final descriptive statistics to be reported are the correlations of Verbal IQ with the performance subtests and Performance IQ with the verbal subtests. The correlations are presented to show that Performance IQ and Verbal IQ correlate fairly well with the subtests for which they would serve as independent internal criteria (Tables 12 and 13).

The correlations, which range from a low of .438 to a high of .708, can be considered moderate in both groups. In both cases the average white correlations were somewhat higher than those of the blacks. This is not atypical when black and white correlations are compared (Jensen, 1980).

Table 11

Reliabilities of Subtests and Summary Scores for Blacks and Whites.

	Black	White
Information	.83	.87
Comprehension	.75	.79
Similarities	.80	.79
Arithmetic	.76	.82
Vocabulary	.90	.92
Picture Completion	.80	.81
Block Design	.79	.82
Picture Arrangement	.61	.68
Object Assembly	.65	.76
Average	.77	.81
Performance Scale	.87	.90
Verbal Scale	.95	.96
Full Scale	.93	.95

Table 12

Correlations of Performance IQ with Five Verbal Subtests.

Group	Information	Comprehension	Similarities	Arithmetic	Vocabulary	Average
White	.612	.568	.581	.647	.594	.600
Black	.456	.566	.538	.550	.547	.531
Difference	.156	.002	.043	.097	.047	.069

Table 13

Correlations of Verbal IQ with Four Performance Subtests.

Group	Picture Completion	Block Design	Picture Arrangement	Object Assembly	Average
White	.644	.659	.708	.573	.646
Black	.588	.533	.590	.438	.537
Difference	.056	.126	.118	.135	.109

Stage One - Factor Analyses

The first stage of the data analysis involved an investigation of the factorial structure of the subtest covariance matrices. (Correlation matrices were actually used for the individual group analysis.)

The investigation began with a principal components analysis followed by a varimax rotation, since this is the procedure most often followed in exploratory factor analytic studies of the Wechsler scales. The "eigenvalue greater than one" criterion was initially employed for determining the number of significant components. Using this technique two significant components were found in both groups. Table 14 presents the factor loadings and eigenvalues for each group.

A visual analysis of the rotated factor plots revealed considerable scatter among the variables in the first quadrant. Five subtests in each group did not appear to load highly on any one factor. Therefore, if two factors were present they would not be easily interpreted in the usual manner.

Following this the EFAP II computer program (Joreskog and Sorbom, 1978) was used to perform exploratory maximum likelihood factor analyses of two, three, and four factor solutions. Since both this procedure and confirmatory factor analysis make use of maximum likelihood methods, it was thought that the findings might prove more helpful for building confirmatory factor analytic models than were the results of the principal components analyses.

It was found that models with either two or three factors could not account for the correlations; however, a model with four factors

Table 14

Factor Loadings and Eigenvalues Resulting from the Principal Components Analyses.

	White		Black	
Subtest	1	2	1	2
Information	736	354	758	163
Comprehension	787	280	782	296
Similarities	624	389	628	342
Digit Span	272	574	352	427
Arithmetic	538	554	586	393
Vocabulary	921	244	875	206
Picture Completion	399	680	367	687
Block Design	382	751	259	795
Picture Arrangement	441	662	470	449
Object Assembly	388	591	138	804
Digit Symbol	071	533	381	290
Eigenvalues	6.175	1.166	5.487	1.298

Note: Decimal points left out.

fit very well in both groups. The results of the goodness of fit tests are in Table 15.

The results of the maximum likelihood analyses suggest that the factor structure is, in both groups, much more complex than had been indicated by the principal components analyses.

Based both on the information gained from the exploratory analyses and on theoretical considerations, a series of confirmatory factor analytic models were then tested for two, three, four, and five factor solutions. After each computer program was run the overall fit of the model, and the values of the various parameters, residuals and first derivatives, were analyzed. Once the major factors (other than the general factor) had been identified in each group, subtest factor loadings were set to zero on one factor at a time. The order in which loeadings were set to zero was usually, but not always, as a result of a small ratio of the factor loading to its standard error. The models finally accepted as best fitting in both groups are presented in Table 16. All of the estimates of factor loadings are more than two standard errors different from zero. Three possible models will be presented for the black group. The first one presented is thought to be the most theoretically acceptable, and hence it is the model that was used in the simultaneous group analysis. In addition, it is this model that was interpreted for the black group.

Table 15

Results of Exploratory Maximum Likelihood Factor Analyses for Two, Three, and Four Factor Solutions.

Group	Number of Factors	Chi-Square	Degrees of Freedom	Probability Level
Black	2	98.8743	34	.000
	3	47.9455	25	.004
	4	19.1111	17	.322
White	2	118.7723	34	.000
	3	43.6606	25	.012
	4	17.3505	17	.431

Table 16

Factor Loadings for Models Accepted as Best Fitting in Both Groups.

	White			
Subtests	1	2	3	4
Information	769	592	295	00*
Comprehension	638	218	00*	549
Similarities	635	00*	00*	460
Digit Span	649	-295	00*	00*
Arithmetic	866	00*	00*	00*
Vocabulary	735	327	00*	467
Picture Completion	666	00*	384	00*
Block Design	649	00*	601	00*
Picture Arrangement	744	00*	272	00*
Object Assembly	476	00*	855	00*
Digit Symbol	469	-328	00*	00*

$X^2 = 22.20$

df = 27

p = .73

Note: Decimal points are left out.

* Factor loading was set to zero.

Table 16 (continued)

Factor Loadings for Models Accepted as Best Fitting in Both Groups.

	Black (Model 1)				
Subtests	1	2	3	4	5
Information	973	00*	00*	00*	00*
Comprehension	638	364	00*	00*	276
Similarities	479	00*	00*	00*	00*
Digit Span	288	00*	00*	585	00*
Arithmetic	548	00*	00*	541	00*
Vocabulary	793	474	00*	00*	00*
Picture Completion	398	201	594	00*	00*
Block Design	383	00*	771	00*	00*
Picture Arrangement	436	00*	289	00*	312
Object Assembly	245	00*	775	00*	00*
Digit Symbol	313	184	00*	271	00*

$X^2 = 29.02$

df = 24

p = .22

Note: Decimal points are left out

* Factor loading was set to zero

Table 16 (continued)

Factor Loadings for Models Accepted as Best Fitting in Both Groups.

	Black (Model 2)				
Subtests	1	2	3	4	5
Information	00*	826	00*	00*	00*
Comprehension	175	494	00*	00*	397
Similarities	225	00*	00*	00*	827
Digit Span	186	00*	00*	571	00*
Arithmetic	00*	00*	00*	811	00*
Vocabulary	134	942	00*	00*	00*
Picture Completion	415	226	576	00*	00*
Block Design	00*	00*	953	00*	00*
Picture Arrangement	00*	00*	307	00*	470
Object Assembly	347	00*	721	00*	00*
Digit Symbol	204	00*	00*	478	00*

$X^2 = 31.28$

df = 28

p = .30

Note: Decimal points are left out.

* Factor loading was set to zero.

Table 16 (continued)

Factor Loadings for Models Accepted as Best Fitting in Both Groups

	Black (Model 3)				
Subtests	1	2	3	4	5
Information	-193	826	00*	00*	00*
Comprehension	00*	500	00*	00*	407
Similarities	00*	00*	00*	00*	855
Digit Span	228	00*	00*	582	00*
Arithmetic	00*	00*	00*	818	00*
Vocabulary	00*	942	00*	00*	00*
Picture Completion	360	262	568	00*	00*
Block Design	00*	00*	951	00*	00*
Picture Arrangement	00*	00*	311	00*	464
Object Assembly	347	00*	730	00*	00*
Digit Symbol	208	00*	00*	490	00*

$X^2 = 27.07$

df = 26

p = .41

Note: Decimal points are left out.

* Factor loading was set to zero.

Composition of Factors

White Group. The first factor is a general factor on which all subtests load at least moderately (from .464 to .866). This factor may represent Spearman's "g" factor or Global Intelligence factor.

The second factor is a bipolar factor. The highest loading is on the Information subtest, with smaller positive loadings on the Comprehensive, and Vocabulary subtests, and negative loadings on the Digit Span and Digit Symbol subtests. Since the two subtests that load negatively on this factor include task requirements that are usually being encountered for the first time during testing, this factor may represent a distinction between use of information and strategies acquired in the past, and those first encountered at the time of the test. It is therefore termed "Acquisition of Information and Strategies."

The third factor is largely consistent with the traditional Perceptual Organization factor. The highest loadings are for the Object Assembly and Block Design subtests. The loadings of Picture Completion and Picture Arrangement are also significant. The one surprise is the loading of the Information subtest on this factor. This may indicate that perceptual organization has more than minimal effect on performance on this subtest.

The fourth factor seems to be a factor of Verbal Fluency. Comprehension, Similarities and Vocabulary load about equally on this factor. In all of these subtests correct answers are scored in one of two ways. One point is given for an answer that is only a specific or concrete instance of a correct answer, and two points are given

for an answer that is a higher level generalization. Usually, the more verbally fluent subjects are, the greater the likelihood that they will receive two point responses.

Black Group (Model 1). The first factor appears to be a General Information factor. The Information subtest loads very highly on this factor, and the Vocabulary and Comprehension subtests have substantial loadings on it. These three subtests are the most information- and content- oriented of the WAIS subtests, and are the most dependent on past learning. The performance subtests, which are much less dependent on past learning, have smaller loadings on this factor.

Factor two is not easily interpreted. The Vocabulary and Comprehensive subtests have the highest loadings on this factor. Picture Completion and Digit Symbol have small loadings on it. All of these subtests load significantly, but to varying degrees, on the General Information factor. Since the Digit Symbol subtest is a rote copying test that requires the motivation to work quickly, this factor could be viewed as "Motivation to Retrieve and Use Information."

Factor three more clearly represents the traditional Perceptual Organization factor in the black sample than in the white sample. Not only are the relevant subtest factor loadings higher but the Information subtest does not load significantly on this factor in the black group.

Factor four is consistent with the Freedom from Distractability factor often found, especially in studies involving non-normal groups. The fact that this factor was present in the black group but not in the white is consistent with interpretations to be made in the discussion section concerning possible selection factors.

Factor five is thought to be similar to the Verbal Fluency factor in the white group. It can be termed "Ability to Discern Relationships." The Similarities subtest loads very highly on this factor and the Comprehension and Picture Arrangement subtests have small but significant loadings on it. A requirement of all these subtests is the ability to see relationships among objects and or concepts. For example, Item 10 on the Comprehension subtest asks "Why are people who are born deaf usually unable to talk?" A correct answer to this item requires one to understand the relationship between hearing and the development of speech. The fact that Picture Arrangement loads significantly on this factor makes it different than the Verbal Fluency factor in the white group. The Picture Arrangement subtest is a nonverbal subtest in which subjects arrange pictures in a correct sequence by determining the relationship among the pictures. The last few items require an understanding of the social context of the sequence or story before one can determine the correct relationship among the pictures. Subjects often verbalize the sequence to themselves when answering these items, so verbal fluency may aid performance.

Once these two models had been identified the question then was whether the differences between them were great enough to suggest differential validity, and thus potential bias, when the test is used as a predictor of ability.

In an effort to answer this question, a series of simultaneous factor analytic models (Joreskog and Sorbom, 1970) was tested in both groups. The two groups were compared on the best fitting model in each. The simultaneous factor analytic models tested involved

invariance of the subtest factor loadings (FL), factor loadings and factor convariances (fc), and factor loadings, factor covariances and error variances (EV). The results of the goodness of fit tests are in Table 17.

None of the models tested fit the observed data near the .05 level. Therefore, the hypothesis stated earlier of equality of subtest factor loadings was not supported by the data. In a further attempt to explore the similarities and differences between the two groups the next models tested involved whether or not the above matrices had the same factor pattern in the samples. For example, if the subtests that load significantly on each factor in one group also load significantly (but to varying degrees) on the same factors in the second group, then the subtest factor loading matrices can be said to have the same pattern. It was found that when the groups were compared using the pattern found in the white group, there was a relatively poor but not significant fit statistic ($X^2(54) = 71.82$, $p = .053$).

Although significant differences in factor structure were found between the white and black samples, some amount of difference could be expected due to selection and cultural factors. It was of some interest whether these differences in factor structure were greater than those found between the present white sample and a normal white sample. This comparison was made because differences between the factor structures of the white and black groups may be due to factors other than those indicative of bias against one racial group. If the factor structures of the white and black samples were much more alike than the factor structures of the present white sample and a normal

Table 17

Results of the Simultaneous Factor Analytic Models Tested.

	Invariant (FL)			Invariant (FL, FC)			Invariant (FL, FC, EV)		
Model	Chi-Square	df	p	Chi-Square	df	p	Chi-Square	df	p
White	221.17	82	.00	158.77	82	.00	179.22	93	.00
Black	192.36	79	.00	162.56	79	.00	182.68	90	.00

white sample, then the differences in factor structure between the present white and black samples may not be due to bias directed against one group.

The normal white sample chosen was the WAIS standardization sample for 18-19 year olds (Wechsler, 1955). This sample of 200 subjects included whites and nonwhites in the ratios found in the 1950 census. The majority (approximately 90%) of this sample was white and for purposes of comparison it can very nearly be considered a normal white sample.

An exploratory maximum likelihood factor analysis suggested that two factors could adequately account for the correlations in the standardization sample ($X^2(34) = 29.957$, $p = .666$). This analysis was followed by testing a confirmatory factor analysis model in which the six verbal subtests loaded on one factor only and the five performance subtests loaded on only the other. This model did not adequately fit the data ($X^2(43) = 47.823$, $p = .00$). However, when the Digit Symbol subtest was allowed to load on both factors, the model fit the data well ($X^2(42) = 47.823$, $p = .248$). The estimates of factor loadings are presented in Table 18.

The Digit Symbol subtest loads predominantly on the first or Verbal Comprehension factor. Its loading on the second, or Perceptual Organization factor, is only slightly larger than its standard error and thus may indicate that the Verbal Comprehension factor could be more complex than is usually thought. This would certainly be consistent with the findings of the present study.

Unfortunately the subtest covariance matrix for the standardization sample was not available, and hence, simultaneous

Table 18

Estimate of Factor Loadings for WAIS Standardization Sample of 18-19 Year Olds.

Subtest	1	2
Information	.898	00*
Comprehension	.779	00*
Arithmetic	.726	00*
Similarities	.851	00*
Digit Span	.644	00*
Vocabulary	.903	00*
Digit Symbol	.609	.119
Picture Completion	00*	.825
Block Design	00*	.836
Picture Arrangement	00*	.739
Object Assembly	00*	.774

Note: * are fixed.

group factor analyses could not be performed. The differences however, between the best fitting models in the white and black groups, appear to be much less than that between the present white sample and the normal white sample. Therefore, the finding of significant differences in factor structure between the present samples, while meaningful for the interpretation of WAIS scores in this and similar samples, cannot necessarily be taken as evidence of bias against the black sample.

Stage Two - Log Linear Model Analyses

Introduction. The second stage of the data analysis involved an analysis of the items to determine if any differential patterns of responding were present. Of the 153 items in the nine subtests evaluated, 98 (65 percent) were suitable for statistical analysis. In general, the remaining 35 percent of the items contained either too few correct or incorrect responses. The average difficulty level of these items (averaging over both groups) was either greater than .94 or less than .06. For the most part the difficulty levels of these extreme items were similar in the two groups.

The fact that so many items could not be evaluated statistically is due largely to the way the test was constructed. The first several items in each subtest were purposely made very easy, and the last several items very hard. Therefore, it is mostly the middle two-thirds of the items that were evaluated.

Due to greater variability, it was easier to fill the required cell frequencies with the independent internal criterion of ability than with the internal criterion. Therefore, fewer cells had to be collapsed with adjacent cells, and a number of items were evaluated with more intervals (and some with more score categories) when the independent internal criterion was used.

The analyses will begin by presenting a table of the means and standard deviations of the scaled scores for both groups. Scaled scores have a mean of 10 and a standard deviation of 3 in the standardization sample. Following this, summary statistics will be reported. The actual statistical results and the probability level of the log-linear models which suggested a bias are presented in

Appendix C. In the case of nonconstant bias, none of the models evaluated for a particular item adequately fit the data. In these instances the models to be presented are the most restrictive models that still suggest a constant bias, i.e. the models containing all possible two factor interactions. Detailed analyses of the items identified as biased are in Appendix D. (Table 19)

Table 19

Mean and Standard Deviation of Subtest Scaled Scores for Whites and Blacks.

	Whites		Blacks	
Subtest	$\bar{X}$	S.D.	$\bar{X}$	S.D.
Information	6.31	2.17	5.47	1.75
Comprehension	7.37	2.67	6.71	2.27
Similarities	8.59	2.52	8.18	2.67
Digit Span	7.74	2.92	7.48	2.82
Arithmetic	6.33	2.45	6.16	1.94
Vocabulary	7.19	2.08	6.19	1.94
Picture Completion	8.24	2.52	7.71	2.21
Block Design	8.41	3.12	7.57	2.74
Picture Arrangement	8.36	2.74	8.31	2.49
Object Assembly	8.87	3.03	7.77	2.60
Digit Symbol	8.14	2.12	7.92	1.83

Summary of the Log Linear Model Analyses

A total of 58 items were identified as biased by at least one criterion. Of these, 24 (41 percent) were identified as biased by both criteria. The direction of bias significantly favored the same group in 15 of 24 (63 percent) of the commonly identified items.

Twenty-eight percent of the items in the nine subtests were identified as biased by the internal criterion. Thirty-two percent of the items were identified as biased by the independent internal criterion. Table 20 presents the number of biased items favoring each group, for each criterion of ability.

Chi-Square tests were performed to test the assumption that biased items would be equally likely to favor each group. Yates correction for continuity was used with all chi-square tests having one degree of freedom. When the groups were compared on the internal criterion no significant differences were found. When they were compared on the independent internal criterion whites were favored on the total of the nine subtests ($X^2(1) = 20.50$, $p < .01$), and on both the verbal section ($X^2(1) = 13.66$, $p < .01$) and performance section ($X^2(1) = 6.71$, $p < .01$).

Evaluation of bonus point responses. Of the six items containing at least the correct-without-bonus and correct-plus-bonus-point score categories evaluated by the internal criterion, four were identified as biased (Picture Arrangement (PA) 7, Block Design (BD) 7, Object Assembly (OA) 1 and 2). All four of these favored the white group. Seven of eight such items evaluated by the independent internal criterion were found biased. All seven (PA 7, BD 7, 8 and 9; OA 1, 2 and 3) favored the white group. All but one item (OA 1) were

Table 20

Number of Biased Items Favoring White and Black Groups for Each Criterion of Ability, for the Total Nine Subtests, and Verbal and Performance Sections.

	Internal Criterion		Independent Internal Criterion	
Subtests	Whites Favored	Blacks Favored	Whites Favored	Blacks Favored
Total	14	6	26	2
Verbal	8	4	19	2
Performance	6	2	7	0

reanalyzed with the two correct score categories collapsed together. OA 1 was a very easy item, so the original log-linear analysis involved only the correct-without-bonus and correct-plus-bonus categories. (For purposes of the high quality analysis described earlier, the incorrect responses were combined with the correct-without-bonus point responses.) If the bonus point responses were a major contributing factor to the statistical bias, then when the two correct score categories were collapsed together fewer items should be found biased. The results of the reanalysis appear to bear this hypothesis out. Two of three items reanalyzed by the internal criterion (PA 7 and OA 2) were still found biased, but only two of six items reanalyzed by the independent internal criterion (BD 8 and OA 2) remained biased.

Next, the individual ability intervals were evaluated. With the internal criterion there was a total of 15 ability intervals evaluated for the six items. Whites were significantly favored in four of these and blacks in none. Although there appeared to be a tendency for whites to be favored in more intervals than blacks, the absolute number was too small to be tested for significance. When the independent internal criterion was used there was a total of 34 ability intervals for the eight items evaluated. Whites were significantly favored in ten of the intervals and blacks in none. This discrepancy proved to be highly significant ($X^2(1) = 9.8$, $p < .01$).

Footnote

[1]In the models (A) stands for ability, (R) for race, and (C) for score category. The (RC) term represents an interaction of race and score category. It is this term that suggests a bias. All but a few of the models reported contain an (AR term. This term represents the fact that there are differing proportions of blacks and whites at each ability level. When the (AR) term was not present it was always the case that some ability intervals had to be collapsed into adjacent intervals.

Discussion

Selection Factors

In most studies comparing blacks and whites on intelligence tests, whites score an average of about 15 points higher than blacks. Male and female scores are usually very close. In the present study whites scored an average of only 3.5 points higher than blacks, and in both groups the male average was approximately 5 points higher than the female average. In addition, whites had IQ scores that were about 10 points lower than that found in normal white samples, and the distribution of IQ scores was skewed, with an overabundance of scores more than one standard deviation below the mean. In the black group the average IQ was close to that found in normal black samples, and the distribution of IQ scores closely approximated a normal distribution.

One possible explanation for these findings is that blacks are more likely than whites to be referred to school psychologists for nonacademic (behavioral) reasons, and males are more likely than females to be referred for nonacademic reasons. This interpretation is based on past findings of a moderate degree of relationship between academic achievement and intelligence, and little relationship between behavior problems and intelligence.

Additional evidence for the hypothesis of differential selection factors can be seen in the extent that the two samples load on the Freedom from Distractability factor. This factor is pronounced in the black group, but absent in the white group. Students referred for nonacademic or behavioral reasons are usually much more easily

distracted from the tasks before them than students referred mainly for academic reasons.

Evidence of Bias Against One Group

The subtest factor structures were found to differ significantly in the two groups. However, the differences in factor structure between these two samples were considerably smaller than the differences between either of these samples and a normal sample. In addition, when the factors of differential selection and differential country of birth (estimated at 15 percent more blacks than whites born outside the United States) are taken into consideration the differences in factor structure appear more understandable. Instead of focusing on group differences in factor structure, perhaps more attention should be paid to the many similarities. The Perceptual Organization factor was largely the same in both samples, and the traditional Verbal Comprehension factor appears to divide into two specific factors in both groups. While one of these, an informational factor, is understandable for the black group because an estimated 28 percent of the black sample was born outside the United States, it is somewhat surprising that an informational factor "Acquisition of Information" was also present in the white group. A possible reason for this is that a number of the items of the WAIS, which was developed in the 1950's may be outdated for an urban population in the 1980's.

The results of the item bias study do not conclusively support an argument of bias against one group. When the internal criterion was used the direction of bias favored no one group. However, when the independent internal criterion was used the direction of bias

favored whites. The fact that the two criteria gave different results suggests that different interpretations must be made to the "bias" they identify. This will be discussed when the two procedures are compared.

Internal Validity of the WAIS for Blacks

While a number of studies have reported on aspects of internal validity (largely factor analyses) of the WAIS for whites, there is a lack of such studies for blacks. If one can assume that the internal validity of the WAIS for whites has been adequately established, then one means of evaluating the internal validity of the test for blacks is by comparing the groups on various indices.

The first indices to be examined are the subtest and summary score reliabilities. Table 11 reveals that while reliabilities tended to be somewhat higher for whites, they were very close in all but two cases, Picture Arrangement and Object Assembly. These two subtests proved to be the least reliable in both groups. In addition, it is interesting that the greatest discrepancy in reliabilities was on the subtest (Object Assembly) that gives the most weight to bonus point responses.

Most past attempts to establish the internal validity of intelligence tests have relied on factor analysis to show that the test is measuring the factors that it is thought to measure (Jensen, 1980). The construction of the Wechsler scales was based on a hierarchical theory of intelligence. They were hypothesized to measure a global or general factor of intelligence and two specific group factors. One of these specific factors was a verbal factor (Verbal Comprehension), the other was a performance factor

(Perceptual Organization). In the present study the groups were found to be highly similar on the Perceptual Organization factor and to have both similarities and differences in the verbal factors found. Where the groups were most discrepant was in their loadings on the general factor. It seems that for these samples the WAIS is a fairly good measure of global intelligence for whites but not for blacks. For blacks the WAIS appears to measure mostly specific factors. Since one of the main reasons that the WAIS is administered to students referred to school psychologists is to obtain a global estimate of intelligence, this finding forces one to question such estimates for the present sample of blacks. If the present findings were to be found in a study with a more representative sample that does not suffer from the selection and cultural factors found in the present study, it could have major implications for interpreting black scores in other samples.

Evaluation of Differences on Performance Subtests

While critics of intelligence tests such as the Wechsler scales have been able to build a number of arguments as to why whites outscore blacks on the verbal subtests, they have had much more difficulty explaining the fact that whites equally outscore blacks on the nonverbal performance subtests.

When critics do attempt to explain the differences on the performance subtests, the main argument given is that whites may have had more exposure to the kinds of tasks involved. An alternative interpretation has been put forth by Sattler (1979) among others: Blacks may not be as motivated as whites for taking these tests. It may be that whites value the payoff (educational, occupational etc.)

of achieving high test scores more than do blacks, and this could account for the greater motivation.

An attempt was made to evaluate this hypothesis in the present study. Ten of twenty two items in three performance subtests allow bonus points for correctly answering an item within certain time limits. The speed with which one answers an item is thought to be due to two factors--ability and motivation. The more highly motivated people are, the faster they will try to complete the task, if they know that the payoff for rapid performance is a higher score.

When the bonus point responses were collapsed into the correct-response-only category, many fewer items were found biased than had been. This suggests that an important discriminating factor between the two group performances was not the ability to answer the items correctly, but rather how fast they were answered. Further, when the independent internal criterion was used to equate the groups on ability, whites were found to have achieved significantly more bonus point responses than blacks in a significant number of ability intervals. This finding is seen as supportive of the hypothesis that the whites in the present study are more highly motivated to achieve higher scores than are the blacks. This finding is strengthened if one can assume that test critics are correct and black ability is actually being underestimated on the verbal subtests. The greatest discrepancy between the groups as to bonus point responses was found when Verbal IQ was used as the independent internal criterion. If the independent internal criterion actually underestimated black ability, then when the groups were equated on it blacks were actually

of higher ability than whites, and therefore could be expected to answer the items more easily.

While these findings appear to confirm the hypothesis that the groups may differ on the motivation to achieve higher scores, it must be remembered that, due to the width of the ability intervals, they were only roughly equated on ability. The alternative hypothesis that the findings were due to the groups differing on ability cannot be totally ruled out. The mean Verbal IQ score for the white group was higher than the mean Verbal IQ score of blacks in intervals 1, 2, 4, and 5. The mean for the black sample was higher in interval 3. Tests of significance indicated that group means differed significantly in only interval 4 ($t(108) = 2.87$, $p < .01$). Only two of ten ability intervals for which whites achieved significantly more bonus point responses than blacks were from interval 4. This suggests that whites achieved more bonus point responses than blacks even when group abilities did not differ significantly.

An alternative explanation, which is related to motivation, that may account for the finding that whites achieved more bonus point responses than did blacks can be found in Hall (1973). Hall, a social anthropologist, studied time orientation in different cultures. He concludes that the meaning assigned to the concept of time varies greatly among cultures. According to Hall, Anglos attach an importance to time that is not found in many other cultures. Therefore, the idea of completing a task within strict time limits may be more familiar to white culture than to black.

Comparison of Internal and Independent Internal Criterion of Ability

The benefits of using a criterion of ability that is external to

the items being evaluated for bias rather than the more usual internal criterion have been discussed (pp. 40, 55). The major benefit is that the use of such a criterion avoids the logical confounding inherent in the use of a criterion (internal) that is dependent on the scores of the items to be evaluated for bias. Shepard, Camilli and Averill (1981) attempted a comparison of the two procedures. Unfortunately, pronounced ceiling effects on their external criterion seriously mitigated their findings and any possible conclusions.

The major finding of the present study is that the two procedures appear to be identifying different aspects of "bias" and therefore are not necessarily interchangeable. The fact that only forty one percent of the items identified as biased were common to the two criterion of ability points up a problem with item bias detection procedures. The definition of a "biased item" is not independent of the specific criterion of ability. If, in the present study, the criterion of ability had been a criterion external to the test itself, such as teacher estimates of intellectual ability, then a totally different set of items may have been identified as biased.

When the internal criterion is used the bias that is identified can be considered ralative bias. Each subtest item is evaluated as to how discrepant it is from a combination of the other items in the subtest. This type of bias can lead to erroneous conclusions. For example, consider the case of a ten item test in which nine items are equally biased to a considerable extent against the same group, and the tenth item is only slightly biased against that group. If an internal criterion of ability such as total test score were used when evaluating the items for bias then only the tenth item would be

so identified. Further, the direction of bias would favor the group which the item is actually biased against. Due to such possibilities this criterion (internal) is considered to be most useful when the great majority of items in a test are unbiased.

On the other hand, if an independent internal or external criterion of ability could be found that correlates highly with the ability measured by the hypothetical ten item test, and if the independent internal or external criterion can provide an unbiased estimate of ability, then the likelihood is that the nine biased items would be correctly identified and the tenth item would not be found to favor the group that it is slightly biased against. In the case of an independent internal or external criterion each item is being evaluated independently of the other items in the same test or subtest. Item bias therefore indicates that two groups which have been equated on a relevant independent internal or external criterion of the ability being measured by the test do not have the same probability of success on the items.

Another way in which the present two criteria were found to differ is that the findings when the independent internal criterion of ability was used appear to be more consistent with the confirmatory factor analyses than when the internal criterion was used. The confirmatory factor analytic study had suggested that the greatest differences between the groups were on the verbal factor(s) rather than the performance factor. When the independent internal criterion was used, the two verbal subtests Arithmetic and Similarities) for which blacks were significantly favored in more ability intervals than whites were the same ones which loaded on

different factors in the two groups. Arithmetic loaded on the general factor in the white group, but predominantly on a specific factor in the black group. Similarities loaded on both the general factor and the Verbal Fluency factor in the white group, but predominantly on the Ability to Discern Relationships factor in the black group.

When using an independent internal or external criterion the main concerns are that it is actually a good indicator of the ability measured by the test, and also that it can be shown to be unbiased (or at least less biased than the internal criterion). In the present study Performance IQ was claimed on logical grounds to be a less biased criterion of the ability being measured by the verbal subtests than was the total score on those subtests. The argument for this was that the content of the verbal items causes most of the criticisms of the Wechsler scales as biased against blacks. However, since approximately 25% of the performance subtest items allow bonus points for rapid performance, and since the blacks in the present study were found to have a significantly lower probability than whites of achieving bonus points, it is questionable whether Performance IQ provided a less biased criterion for the present sample.

Conclusions

The statistical findings of the present study are not unambiguous since they are affected both by possible differential selection factors, and by probable differences in the amount of time that the groups have been exposed to American culture. They nevertheless add to our methodological and theoretical knowledge in the area of test and item bias.

Methodologically, the study has shown some of the advantages of confirmatory factor analysis for developing a greater understanding of the similarities and differences in the factor structure of test scores for two groups. Further, an independent internal criterion of ability was used in addition to the internal criterion traditionally used when investigating item bias. The independent internal criterion avoids the logical confounding inherent in the use of an internal criterion. The findings observed when this procedure was used were consistent with certain findings of the confirmatory factor analyses. Therefore, the use of an independent internal or external criterion of ability is warranted in further studies of item bias.

The present study is unique in that past investigations of bias in the Wechsler scales have used young children. If, as is asserted by critics such as Williams, blacks and whites differentially develop certain cognitive skills and ways of viewing things as a result of the differing environmental and cultural experiences to which they have been exposed, then one should find that the differences between the groups becomes greater with age. While differences between the groups were found in the present study, there were many similarities between them. The similarities would probably be even greater if the selection factors and cultural factors related to country of birth could be controlled. It is concluded that the WAIS is internally valid in comparable ways for blacks and whites. However, for this particular sample of blacks, the WAIS seems to be a better measure of specific factors than of any general factor of intelligence. If future studies which are more representative result in similar findings for the general and specific factors, then the

conclusion concerning the internal validity of the WAIS for blacks may have to be altered.

The findings of the present study do not confirm the arguements of test critics who have stated that blacks score more poorly than whites on the Wechsler scales because the test content is biased against them. An alternative explanation was put forth that may account for part of the black-white differences in IQ scores, found in the present sample: Blacks may be less motivated than whites to perform to the full extent of their abilities.

If motivation is a factor in blacks lower scores on performance subtests that allow bonus point responses, then it can be expected that this discrepancy will be more pronounced in at least one of the performance subtests of the newly revised WAIS (Wechsler, 1981). In the older version, four of nine items in the Block Design subtest allow bonus points. In the newly revised version, seven of nine items allow bonus points, so the effect should be stronger in the new version.

Recommendation for Test Interpretation

1. The WAIS was found to be a fairly good estimator of global intelligence in the white sample, but not so for the black sample. For the present sample of urban blacks the WAIS was found to be a better measure of specific factors than of any global factor of intelligence. Therefore, if Full Scale IQ scores are treated as global estimates of intelligence when decisions, expecially those involving CRMD placements are made, there is a possibility that more blacks than whites are being misplaced.

Recommendations for Future Research

1. The findings of this study are limited both by apparent differential selection factors and by probably differences in country of birth, and hence cultural factors. What is now needed is a study on a larger and more representative sample of blacks and whites. There exists a large data base that could be tapped. Most School and Clinical Psychology programs routinely require their graduate students to administer the WAIS. Therefore, what is needed is a way to organize the data collection and interpretation.
2. The hypothesis of differential motivation for taking intelligence tests should be more fully explored and its effects more clearly delineated. For example, if one could control for motivation by using teacher estimates (ratings) of student motivation as a covariate when evaluating the black-white difference on Performance IQ, or on specific performance subtests, would the groups still differ significantly.

Recommendations for Policy

1. While the school psychology literature has for years recommended that individual intelligence test data be but one component in the assessment of mental retardation, administrators do not always follow this advice. Unfortunately, too many decisions, such as placement in certain vocational training programs, are still made largely on the basis of an IQ score. The findings of the present study reinforce the importance of including data from other sources when making a diagnosis of mental retardation, especially in minority groups.
2. If future investigations performed on similar samples of blacks and whites, confirm the finding that the groups differ as to the motivation needed to achieve higher test scores, then it might prove beneficial for a system such as the New York City Board of Education to develop local norms for the use of the Block Design and Object Assembly subtests as power tests. The combined score on these two measures, which are the best measures of the Perceptual Organization factor, could then be used as an indication of ability that is not confounded by the motivation to work rapidly.

Appendix A. Description of WAIS Subtests.

Verbal Scale

Information: 29 questions that cover a wide variety of information that adults have presumably had an opportunity to acquire in our culture.

Comprehension: 14 questions in which the subject explains what should be done under certain circumstances, why certain practices are followed, or the meaning of proverbs.

Arithmetic: 14 numerical word problems. The content of the problems progresses from simple addition and subtraction to beginning fractions and percentages. Bonus points are given for successful completion of the last four items within certain time limits.

Similarities: 13 items requiring the subject to tell in what ways two seemingly dissimilar objects or concepts are alike.

Digit Span: In the first part lists of from three to nine digits are to be reproduced. In the second part, lists of from two to eight digits must be reproduced in the reverse order of their presentation.

Vocabulary: 40 words of increasing difficulty are presented to the subject (who also has a list of them in front of him or her). The subject is required to state what the meaning of each word is.

Performance Scale

Digit Symbol: This is a rote copying task in which the subject is given a list of digits and has to fill in the symbol paired with that digit on a key.

Performance Scale (continued)

Picture Completion: 21 cards each containing a picture with some essential part missing are presented. The subject tells what is missing from each picture.

Block Design: 10 items that require the subject to reproduce designs of increasing complexity from multicolored blocks.

Picture Arrangement: 8 items in which the subject must arrange a set of cards in a sequence that is logical and tells a story.

Object Assembly: 4 puzzle tasks of increasing complexity are presented to the subject.

Appendix B.

Subtest Correlation Matrices. Whites N = 182

	Info.	Comp.	Sim.	DN	Arith	Voc.	PC	BD	PA	OA	DS
Information	1.00	.663	.533	.363	.669	.772	.532	.539	.585	.530	.209
Comprehension	.663	1.00	.651	.382	.539	.788	.517	.522	.508	.489	.194
Similarities	.533	.651	1.00	.439	.551	.682	.484	.520	.538	.461	.270
Digit Span	.363	.382	.439	1.00	.575	.381	.457	.494	.513	.363	.393
Arithmetic	.669	.539	.551	.575	1.00	.633	.550	.570	.657	.412	.380
Vocabulary	.772	.788	.682	.381	.633	1.00	.569	.525	.555	.474	.212
Picture Completion	.532	.517	.484	.457	.550	.569	1.00	.674	.589	.605	.457
Block Design	.535	.522	.520	.494	.570	.525	.674	1.00	.636	.769	.366
Picture Arrangement	.585	.508	.538	.513	.657	.555	.589	.636	1.00	.567	.395
Object Assembly	.530	.489	.461	.363	.412	.474	.605	.769	.567	1.00	.231
Digit Symbol	.209	.194	.270	.393	.380	.212	.457	.366	.395	.231	1.00

Appendix B. (continued)

Subtest Correlation Matrices. Blacks N = 267

	Info.	Comp.	Sim.	DN	Arith	Voc.	PC	BD	PA	OA	DS
Information	1.00	.621	.467	.379	.533	.772	.389	.374	.422	.239	.304
Comprehension	.621	1.00	.656	.386	.550	.751	.494	.429	.479	.359	.379
Similarities	.467	.656	1.00	.400	.529	.560	.478	.376	.525	.338	.343
Digit Span	.279	.386	.400	1.00	.479	.369	.408	.400	.385	.379	.256
Arithmetic	.533	.550	.529	.479	1.00	.552	.442	.493	.410	.361	.387
Vocabulary	.773	.751	.560	.369	.552	1.00	.491	.389	.497	.312	.401
Picture Completion	.389	.494	.478	.408	.442	.491	1.00	.645	.461	.624	.342
Block Design	.374	.429	.376	.400	.493	.389	.645	1.00	.512	.689	.291
Picture Arrangement	.422	.479	.525	.385	.410	.497	.461	.512	1.00	.396	.301
Object Assembly	.239	.359	.338	.379	.361	.312	.624	.689	.396	1.00	.308
Digit Symbol	.304	.379	.343	.256	.387	.401	.342	.291	.301	.308	1.00

Appendix C.

Log-Linear Models Accepted as Best Fitting the Data, Which Contained a Term Suggesting a Bias.

(Items identified as biased by the internal criterion.)

Subtest	Item	Log-Linear Model	Chi-Square	Degrees of Freedom	Probability
Information	5	AR,AC,RC	3.80	3	.28
	6	Same	19.59	2	.28
	9	Same	18.61	3	.00
	11	Same	8.27	2	.02
Comprehension	8	AC,RC	3.83	4	.43
	10	AR,AC,RC	11.01	2	.00
	11	Same	.49	2	.78
Similarities	2	Same	1.50	4	.83
	3	Same	1.39	3	.71
	9	Same	17.08	4	.00
	10	Same	6.55	2	.04
Arithmetic	4	Same	6.89	2	.03
	5	Same	12.65	3	.01
	6	Same	9.42	3	.02
	12	AC,RC	2.43	2	.30
Vocabulary	7	AR,AC,RC	12.15	4	.02
	8	Same	5.82	3	.12
	9	AC,RC	1.98	4	.74

Appendix C. (continued)

Subtest	Item	Log-Linear Model	Chi-Square	Degrees of Freedom	Probability
Vocabulary (continued)	11	AR,AC,RC	12.49	4	.01
	12	Same	7.64	3	.01
	13	Same	.65	2	.72
	19	Same	16.36	4	.00
	20	Same	11.16	2	.00
	21	Same	9.35	2	.01
	25	AC,RC	2.58	2	.27
	29	AC,RC	.72	2	.70
Picture Completion	4	AR,AC,RC	15.80	4	.00
	5	Same	4.10	2	.13
	6	Same	12.85	4	.01
	8	Same	18.87	3	.00
	9	Same	12.82	2	.00
	12	Same	9.65	3	.02
	13	AC,RC	8.32	6	.22
	14	AC,RC	9.85	6	.13
	16	AR,AC,RC	8.67	3	.03
	17	Same	10.47	3	.02
	18	Same	8.19	2	.02
Block Design	7	AC,RC	5.88	3	.12

Appendix C. (continued)

Subtest	Item	Log-Linear Model	Chi-Square	Degrees of Freedom	Probability
Picture Arrangement	3	AR,AC,RC	16.63	2	.00
	7	Same	16.88	4	.00
	8	Same	5.86	1	.02
Object Assembly	1	Same	3.15	3	.37
	2	Same	7.95	2	.02

Appendix C (continued)

(Items identified as biased by the independent internal criterion.)

Subtest	Item	Log-Linear Model	Chi-Square	Degrees of Freedom	Probability
Information	5	AR,AC,RC	1.61	4	.81
	9	Same	13.25	4	.01
	10	Same	12.57	4	.01
	11	Same	5.54	4	.24
	12	Same	11.25	4	.02
	15	Same	11.04	4	.03
	16	Same	1.95	2	.38
Comprehension	4	Same	14.36	6	.03
	6	Same	20.53	8	.01
	8	Same	3.88	4	.42
	10	Same	9.98	8	.27
	12	Same	5.59	6	.47
Similarities	3	Same	3.14	4	.53
	5	Same	13.38	4	.01
	7	Same	15.52	8	.00
	9	Same	11.86	4	.02
Arithmetic	5	Same	5.83	4	.21
	8	Same	17.14	4	.00
	9	Same	4.81	4	.31
	10	Same	16.82	4	.00
	12	Same	3.02	2	.22

Appendix C (continued)

Subtest	Item	Log-Linear Model	Chi-Square	Degrees of Freedom	Probability
Vocabulary	7	Same	15.31	6	.02
	8	Same	8.21	4	.08
	9	Same	2.67	3	.45
	10	Same	6.33	3	.45
	11	Same	9.61	6	.14
	13	Same	9.55	6	.14
	14	Same	17.69	6	.01
	17	Same	16.51	4	.00
	18	Same	24.74	4	.00
	20	AR,AC,RC	4.09	4	.39
	24	Same	12.36	4	.01
	25	Same	6.36	2	.04
	29	Same	9.05	3	.03
Picture Completion	5	Same	.51	3	.92
	7	Same	13.21	3	.00
	10	Same	10.18	4	.04
	13	AC,RC	8.08	6	.23
	14	AR,AC,RC	3.59	4	.47
	15	Same	11.20	3	.01
Block Design	2	Same	16.49	4	.00
	7	Same	10.09	6	.12

Appendix C (continued)

Subtest	Item	Log-Linear Model	Chi-Square	Degrees of Freedom	Probability
Block Design (continued)	8	Same	16.90	6	.01
	9	Same	10.81	6	.09
Picture Arrangement	7	Same	6.71	6	.40
Object Assembly	1	AC,RC	13.04	8	.11
	2	AC,RC	12.05	9	.21
	3	AR,AC,RC	5.52	6	.48

Appendix D: Detailed analysis of biased items.

A detailed analysis of the items identified as biased will be presented in tabular form. The tables for each subtest include the items identified as being biased by either the internal or external criterion. If quality comparisons were involved (A), in the tables, stands for incorrect responses compared to both categories of correct responses, and (B) stands for low quality responses (zero and one point responses, or zero and correct without bonus point responses) compared to high quality responses (two point responses, or correct with bonus point responses.). This is followed by noting the criterion that identified the bias, and whether the bias was constant (C) or nonconstant (NC). The direction of bias for the total item is then reported. Probability levels are noted for signed chi-square statistics only when all of the signs are in the same direction. When they are reported, it is noted whether the probability was less than .05 (*) or .01 (**). When they are not reported it can be assumed that the same group was not consistently favored in all intervals, and therefore the signed chi-square statistic is not distributed as a chi-square. In these instances, if the direction of the bias is said to favor one group, the value of the signed chi-square is such that it would be significant at the .05 level had the signed chi-square been distributed as a chi-square. As mentioned previously, this is a conservative estimate. The intervals were then analyzed similarly. Since the itmes were analyzed with varying numbers of ability intervals, the following procedure was followed: Items with an even number of ability intervals had the intervals evenly divided into lower (L) and higher (H) intervals. Items with an odd number of ability intervals

contained an additional category (M) or the middle interval. Items with four or five intervals therefore had two lower and two higher intervals. The lowest interval is L 1 and the highest is H 2.

Table 21

Analysis of the Biased Items in the Information Subtest.

Criterion			Overall Results				Results for Ability Subintervals				
	Identified By		Favored	Signed			Biased	Favored	Signed		
Item	Int.	Ext.	Group	Chi-Square	df	Signif.	Intervals	Group	Chi-Square	df	Signif.
5	C		None	5.39	4		None				
5		C	None	9.34	5		L 2	White	3.84	1	*
6	NC		Black	16.06	3		M	Black	16.60	1	**
9	NC		None	7.81	4		L 1	Black	5.45	1	*
							H 2	White	12.80		**
9		NC	White	15.15	5		H 1	White	13.89	1	**
10		NC	White	12.18	5		L 2	White	8.30	1	**
							H 1	White	5.31	1	*
11	NC		White	13.53	4		H 2	White	12.71	1	**
11		C	White	29.44	5	**	H 1	White	21.09	1	**
							H 2	White	4.47	1	*

Table 21 (continued)

Criterion			Overall Results				Results for Ability Subintervals				
	Identified By		Favored	Signed			Biased	Favored	Signed		
Item	Int.	Ext.	Group	Chi-Square	df	Signif.	Intervals	Group	Chi-Square	df	Signif.
12		NC	White	29.21	5	**	H 1	White	21.09	1	**
							H 2	White	4.28	1	*
15		NC	White	17.72	5		H 1	White	14.51	1	
16		C	White	13.61	3	**	H	White	7.05	1	**

Table 22

Analysis of the Biased Items in the Comprehension Subtest.

Criterion			Overall Results				Results for Ability Subintervals				
	Identified By		Favored	Signed			Biased	Favored	Signed		
Item	Int.	Ext.	Group	Chi-Square	df	Signif.	Intervals	Group	Chi-Square	df	Signif.
4(A)		NC	None	2.13	4		None				
4(B)		NC	None	7.91	4		L 1	Black	9.33	1	**
6(A)		NC	None	8.58	4		L 2	White	5.11	1	*
6(B)		NC	None	3.40	4		H 1	Black	3.98	1	*
8	C		White	9.43	3	*	None				
8(A)		C	None	5.64	3		L	White	5.38	1	*
8(B)		C	White	8.47	3	*	L	White	6.47	1	*
10	NC		White	19.70	3		M	White	11.04	1	**
							H	White	10.31	1	**
10(A)		C	White	22.89	5	**	L 2	White	12.81	1	**
							M	White	6.41	1	*

Table 22 (continued)

Criterion			Overall Results				Results for Ability Subintervals				
	Identified By		Favored	Signed			Biased	Favored	Signed		
Item	Int.	Ext.	Group	Chi-Square	df	Signif.	Intervals	Group	Chi-Square	df	Signif.
10(B)		C	None	10.13	5		L 2	White	10.47	1	**
11(C)	C		None	7.19	3		None				
12(A)	C		None	6.23	4		L 1	White	5.62	1	*
12(B)	C		White	13.23	4	*	L 1	White	4.51	1	*
							H 1	White	6.18	1	*

Table 23

Analysis of the Biased Items in the Similarities Subtest.

Criterion			Overall Results				Results for Ability Subintervals				
	Identified By		Favored	Signed			Biased	Favored	Signed		
Item	Int.	Ext.	Group	Chi-Square	df	Signif.	Intervals	Group	Chi-Square	df	Signif.
2(A)	C		None	3.24	3		None				
2(B)	C		None	7.10	3		H	Black	4.79	1	*
3	C		None	6.66	4		None				
3		C	None	6.68	4		None				
5(A)		NC	None	.30	3		L	White	5.52	1	*
							H	Black	4.18	1	*
5(B)		NC	None	3.01	3		None				
7(A)		NC	None	4.81	5		L 1	Black	4.47	1	*
7(B)		NC	None	1.67	5		L 1	Black	5.23	1	*
							L 2	White	9.61	1	**
9(A)	NC		None	.09	3		None				

Table 23 (continued)

Criterion			Overall Results				Results for Ability Subintervals				
	Identified By		Favored	Signed			Biased	Favored	Signed		
Item	Int.	Ext.	Group	Chi-Square	df	Signif.	Intervals	Group	Chi-Square	df	Signif.
9(B)	NC		None	.72	3		L	Black	3.98	1	*
							M	White	4.69	1	*
9(A)		NC	Black	10.69	3		H	Black	10.39	1	*
9(B)		NC	None	4.71	3		None				
10	NC		White	16.77	3		M	White	8.59	1	**
							H	White	8.42	1	**

Table 24

Analysis of the Biased Items in the Arithmetic Subtest.

Criterion			Overall Results				Results for Ability Subintervals				
	Identified by		Favored	Signed			Biased	Favored	Signed		
Item	Int.	Ext.	Group	Chi-Square	df	Signif.	Intervals	Group	Chi-Square	df	Signif.
4	NC		None	2.58	3		L	White	5.71	1	*
5	NC		Black	10.68	4		L 1	Black	8.49	1	**
							H 1	Black	3.96	1	*
5		C	None	10.81	5		L 1	Black	5.58	1	*
6	NC		White	18.48	4		H 1	White	13.60	1	**
							H 2	White	4.73	1	*
8		NC	None	4.43	5		L 1	Black	5.79	1	*
							L 2	White	5.33	1	*
							M	Black	4.11	1	*
9		C	Black	11.48	5	*	M	Black	4.09	1	*
							H 2	Black	4.32	1	*
10		NC	None	3.74	5		L 1	Black	4.00	1	*
							L 2	White	10.04	1	**

Table 24 (continued)

Criterion			Overall Results				Results for Ability Subintervals				
	Identified by		Favored	Signed			Biased	Favored	Signed		
Item	Int.	Ext.	Group	Chi-Square	df	Signif.	Intervals	Group	Chi-Square	df	Signif.
12	C		None	5.25	2		None				
12		C	None	6.85	3		None				

Table 25

Analysis of the Biased Items in the Vocabulary Subtest.

Criterion			Overall Results				Results for Ability Subintervals				
	Identified By		Favored	Signed			Biased	Favored	Signed		
Item	Int.	Ext.	Group	Chi-Square	df	Signif.	Intervals	Group	Chi-Square	df	Signif.
7(A)		NC	None	5.59	4		L 2	White	4.14	1	*
7(B)		NC	None	.53	4		H 2	Black	4.25	1	*
7(A)	NC		None	2.92	3		H	White	5.18	1	*
7(B)	NC		None	1.33	3		None				
8	C		Black	15.35	4	**	H 1	Black	10.69	1	**
8		C	None	9.95	5		H 2	Black	8.06	1	**
9	C		White	24.01	3	**	L	White	6.06	1	*
							M	White	6.64	1	**
							H	White	11.03	1	**
9		C	White	31.50	4	**	L 1	White	5.61	1	*
							L 2	White	16.64	1	**
							H 1	White	5.13	1	*
							H 2	White	3.88	1	*

Table 25 (continued)

Criterion			Overall Results				Results for Ability Subintervals				
	Identified By		Favored	Signed			Biased	Favored	Signed		
Item	Int.	Ext.	Group	Chi-Square	df	Signif.	Intervals	Group	Chi-Square	df	Signif.
10		C	White	14.53	4	**	L 2	White	11.93	1	**
11(A)	NC		White	15.19	3	**	M	White	3.97	1	*
							H	White	10.94	1	**
11(B)	NC		None	3.31	3		None				
11(A)		C	White	26.50	4	**	L 2	White	20.26	1	**
11(B)		C	White	14.07	4	**	L 2	White	9.84	1	**
12	C		Black	11.03	4	**	H 2	Black	10.19	1	**
13	C		None	6.87	3		L	White	5.47	1	*
13(A)		C	White	10.57	4		L 2	White	6.97	1	**
13(B)		C	None	4.07	4		H 1	Black	5.82	1	*
14(A)		NC	None	7.71	4		H 2	White	5.28	1	*
14(B)		NC	White	14.76	4		H 2	White	12.21	1	**

Table 25 (Continued)

Criterion			Overall Results				Results for Ability Subintervals				
	Identified By		Favored	Signed			Biased	Favored	Signed		
Item	Int.	Ext.	Group	Chi-Square	df	Signif.	Intervals	Group	Chi-Square	df	Signif.
17(A)		NC	None	3.65	3		L	White	4.77	1	*
17(B)		NC	White	9.78	3		L	White	12.17	1	**
18(A)		NC	None	2.28	3		None				
18(B)		NC	White	14.01	3		L	White	15.95	1	**
							M	Black	4.13	1	*
19(A)	NC		None	1.29	3		L	White	5.09	1	*
							M	Black	6.36	1	*
19(B)	NC		None	7.17	3		L	White	8.36	1	**
20(A)		NC	None	6.49	3		None				
20(B)		NC	None	.10	3		None				
20	C		None	2.36	3		M	White	6.65	1	**
							H	Black	4.72	1	*
21	NC		None	4.02	3		H	Black	5.20	1	*

Table 25 (continued)

Criterion			Overall Results				Result for Ability Subintervals				
	Identified By		Favored	Signed			Biased	Favored	Signed		
Item	Int.	Ext.	Group	Chi-Square	df	Signif.	Intervals	Group	Chi-Square	df	Signif.
24(A)		NC	White	11.01	3		L	White	10.29	1	**
24(B)		NC	White	8.70	3		L	White	8.93	1	**
25	C		None	5.87	2		L	White	4.13	1	*
25		NC	White	11.25	3	*	L	White	7.22	1	*
29	C		White	7.53	2	*	L	White	4.26	1	*
29		NC	White	17.91	4	**	L 1	White	10.17	1	**
							H 2	White	6.36	1	*

Table 26

Analysis of the Biased Items in the Picture Completion Subtest.

Criterion			Overall Results				Results for Ability Subintervals				
	Identified By		Favored	Signed			Biased	Favored	Signed		
Item	Int.	Ext.	Group	Chi-Square	df	Signif.	Intervals	Group	Chi-Square	df	Signif.
4	NC		Black	12.45	5		M	Black	11.71	1	**
5	C		Black	8.24	3		L	Black	4.02	1	*
							M	Black	4.23	1	*
5		C	None	4.88	4		None				
6	NC		None	4.09	5		L 1	White	4.39	1	*
							H 1	Black	6.01	1	*
7		NC	None	3.68	4		H 2	White	8.21	1	**
8	NC		None	6.26	4		H 1	White	9.39	1	**
							H 2	Black	3.90	1	*
9	NC		None	5.42	3		H	White	8.82	1	
10		NC	None	6.30	5		L 1	Black	6.69	1	**
12	NC		None	6.19	4		None				

Table 26 (continued)

Criterion			Overall Results				Results for Ability Subintervals				
	Identified By		Favored	Signed			Biased	Favored	Signed		
Item	Int.	Ext.	Group	Chi-Square	df	Signif.	Intervals	Group	Chi-Square	df	Signif.
13	C		White	34.20	4		L 2	White	7.35	1	**
							H 1	White	15.81	1	**
							H 2	White	11.22	1	**
13		C	White	27.58	4		H 1	White	12.81	1	**
							H 2	White	12.88	1	**
14	C		White	12.01	4		L 1	White	6.73	1	**
							L 2	White	5.32	1	*
14		C	None	10.23	5		L 2	White	6.35	1	*
15		NC	None	9.30	4		H 1	White	5.04	1	*
							H 2	White	5.19	1	*
16	NC		None	6.02	4		L 1	Black	4.27	1	*
17	NC		None	8.31	4		L 1	Black	8.17	1	**
18	NC		None	5.07	3		L	Black	5.39	1	*

Table 27

Analysis of the Biased Items in the Block Design Subtest.

Criterion			Overall Results				Results for Ability Subintervals				
	Identified By		Favored	Signed			Biased	Favored	Signed		
Item	Int.	Ext.	Group	Chi-Square	df	Signif.	Intervals	Group	Chi-Square	df	Signif.
2		NC	None	1.13	3		None				
7(A)	NC		None	1.73	2		None				
7(B)	NC		White	20.65	2		H	White	20.65	1	**
7(A)		C	None	1.07	4		None				
7(B)		C	White	14.97	4		L 2	White	11.76	1	**
8(A)		NC	White	12.40	4	*	L 2	White	4.53	1	*
							H 2	White	5.25	1	*
8(B)		NC	White	13.48	4	**	L 1	White	7.22	1	**
							L 2	White	5.86	1	*
9(A)		C	None	8.79	4		L 2	White	7.47	1	**
9(B)		C	White	10.04	4		L 2	White	10.21	1	**

Table 28

Analysis of the Biased Items in the Picture Arrangement Subtest.

Criterion			Overall Results				Results for Ability Subintervals				
	Identified By		Favored	Signed			Biased	Favored	Signed		
Item	Int.	Ext.	Group	Chi-Square	df	Signif.	Intervals	Group	Chi-Square	df	Signif.
3	NC		None	8.61	3		L	Black	13.21	1	**
								White	4.06	1	*
7(A)	NC		None	3.45	3		H	White	5.93	1	*
7(B)	NC		White	11.60	3		H	White	13.41	1	**
7(A)		C	None	4.21	4		None				
7(B)		C	None	4.38	4		H 1	White	4.27	1	*
8	NC		White	8.18	2		L	White	8.52	1	**

Table 29

Analysis of the Biased Items in the Object Assembly Subtest.

Criterion			Overall Results				Results for Ability Subintervals				
	Identified By		Favored	Signed			Biased	Favored	Signed		
Item	Int.	Ext.	Group	Chi-Square	df	Signif.	Intervals	Group	Chi-Square	df	Signif.
1(B)	C		White	14.13	4	**	L 2	White	9.50	1	**
1(B)		C	White	25.20	5	**	L 2	White	6.78	1	**
							H 1	White	6.94	1	**
							H 2	White	8.91	1	**
2(A)	C		White	15.20	2	**	H	White	14.74	1	**
2(B)	C		White	16.91	2	**	H	White	15.69	1	**
2(A)		C	White	29.69	4	**	L 2	White	13.27	1	**
							H 2	White	14.72	1	**
2(B)		C	White	24.51	4	**	L 2	White	14.18	1	**
							H 2	White	8.31	1	**
3(A)		NC	White	9.62	4		H 2	White	7.44	1	**
3(B)		NC	None	3.54	4		None				

References

Alderman, D. L., & Holland, P. W. (1981). Item performance across native language groups on the Test of English as a Foreign Language. TOEFL Research Reports, Report 9: ETS Research Report 81-16. Princeton, N.J.: Educational Testing Service.

American Psychological Association, American Educational Research Association, and National Council on Measurement in Education. (1974). Standards for educational and psychological tests. Washington, D.C.: American Psychological Association.

Anastasi, A. (1968). Psychological testing: 4th edition. New York: McMillan, Inc.

Baker, F. B. (1981). A criticism of Scheuneman's item bias technique. Journal of Educational Measurement, 18, 59-62.

Berk, R. A. (1982). Handbook of methods for detecting test bias. Baltimore: Johns Hopkins University Press, p2.

Birnbaum, A. (1968). Test scores, sufficient statistics and the information structure of tests. In P. M. Lord and M. R. Novick, Statistical theories of mental test scores. Reading, MA.: Addison-Wesley.

Boozer, B. (1978). An alternative to intelligence testing for minority children. Journal of Negro Education, 47 (4), 414-418.

Camilli, G. A. (1979). A critique of the chi-square method for assessing item bias. Unpublished paper, Laboratory of Educational Research, University of Colorado, Boulder.

Cohen, J. (1957). The factorial structure of the WAIS between early childhood and old age. Journal of Consulting Psychology, 21, 283-290.

Cotter, D. E., & Berk, R. A. (April 1981). Item bias in the WISC-R using black, white and hispanic learning disabled children. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles.

Covin, T. M., & Covin, J. N. (August 1976). Comparability of Peabody and WAIS scores among adolescents suspected of being mentally retarded. Psychological Reports, 39 (1), 33-34.

Covin, T. M., & Hatch, G. L. (1977). WISC-R Full Scale mean IQ's for both black and white children aged 6 through 15 and having problems in school. Psychological Reports, 40 (1), 281-282.

Dean, R. S. (1979). Predictive validity of the WISC-R with Mexican American children. Journal of School Psychology, 17 (1), 55-58

Diana et al. vs. California State Board of Education. Civil action No. C-70-37 RFP (N.D. Cal., Jan. 7, 1970; Jan. 18, 1973), consent decree.

Dove, A. (July 15, 1968). Taking the Chittling test. Newsweek.

The Education for all handicapped Children Act of 1975 (Public Law 94-142). (November 29, 1975). In United States Code: Congressional and Administrative News (Vol. 1). St. Paul, MN: West Publishing Co., pp. 773-796.

Goodman, L. A. (1978). Analyzing qualitative/categorical data: Log-linear models and latent structure analysis. Cambridge, MA: ABT.

Green, D. R. (1975). What does it mean to say a test is biased? Education and Urban Society, 8, 33-52.

Gross, A., & Su, W. (1975). Defining a "fair" or "unbiased" selection model: A question of utilities. Journal of Applied Psychology, 60 345-351.

Gutkin, T. B., & Reynolds, C. R. (April 1981). Factorial similarity of the WISC-R for white and black children from the standarization sample. Journal of Educational Psychology, 73 (2), 227-231.

Hall, E. (1973). The silent language. New York: Anchor Books.

Hardy, J. B., Welcher, D. W., Mellits, E. D., & Kagan, J. (1976). Pitfalls in the measurement of intelligence: Are standard intelligence tests valid instruments for measuring the intellectual potential of urban children? Journal of Psychology, 44 (1), 43-51.

Holowinsky, I. Z., & Pascale, P. J. (1972). Performance on selected WISC subtests of subjects referred for psychological evaluation of educational difficulties. Journal of Special Education, 6 (3), 231-235.

Hunt, J. M. (1974). Black genes - white environment: Intelligence tests measure learned performance, not innate capacity. In R. A. Miner (Ed.), Annual editions: Readings in human development 74/75. Guileford, CT: The Dushkin Publishing Group Inc.

Hunter, J. E. (December 1975). A critical analysis of the use of item mean and item-test correlations to determine the presence or absence of content bias in achievement test items. Paper presented at the National Institute of Education Conference on Test Bias, Maryland.

Ironson, G. H. (1982). Use of chi-square and latent trait approaches for detecting item bias. In R. Berk (Ed.), Handbook of methods for detecting test bias. Baltimore: Johns Hopkins University Press.

Ironson, G. H., & Subkoviak, M. J. (1979). A comparison of several methods of assessing item bias. Journal of Educational Measurement, 16, 209-225.

Jensen, A. R. (1980). Bias in mental testing. New York: The Free Press.

Joreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. Psychometrikia, 34, 183-202.

Joreskog, K. G. (1970). A general method for analysis of covariance structures. Biometrika, 57, 239-251.

Joreskog, K. G. (1971). Simultaneous factor analysis in several populations. Psychometrika, 36, 409-426.

Joreskog, K. G., & Sorbom, D. (1979). Advances in factor analysis and structural equation models. Cambridge, MA: ABT Books.

Kaufman, A. S., & Doppelt, J. E. (1976). Analysis of WISC-R standardization data in terms of the stratification variables. Child Development, 47 (1), 165-171.

Larry, P., et al. vs. Wilson Riles, Superintendent of Public Instruction for the State of California, et al., 343 F. Supp. 1306, (N.D. Cal., 1972); Aff'd 502 F. 2d 963 (9th Cir., 1974).

Long, P. A., & Anthony, J. J. (1974). The measurement of mental retardation by a culture specific test. Psychology in the Schools, 11 (3), 310-312.

Lord, F. (1980). Applications of item response theory to practical testing problems. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Marascuilo, L. A., & Slaughter, R. E. (1981). Statistical procedures for identifying possible sources of bias based on chi-square statistics. Journal of Educational Measurement, (Winter), 18 (4), 229-248.

Matarrazo, J. D., & Wiens, A. N. (1977). Black Intelligence Scale of Cultural Homogeneity and Wechsler Adult Intelligence Scale scores of black and white police applicants. Journal of Applied Psychology, 62 (1), 57-63

Matsui, T., Okada, A., & Kakuyama, T. (October 1981). Influence of achievement need on goal setting, performance, and feedback effectiveness. Journal of Applied Psychology, 67 (5), 645-648.

McGaw, B., & Joreskog, K. G. (1971). Factorial invariance of ability measures in groups idffering in intelligence and socioeconomic status. British Journal of Mathematical and Statistical Psychology, 24, 154-168.

McGrevy, D. F., Knouse, S. B., & Thompson, R. A. (March 1974). Relationships among an individual intelligence test and two air force screening and selection tests. U. S. AFHRL Technical Report, No. 74-25, 18.

Mercer, J. R. (1973). Labeling the retarded. Berkeley, CA: University of California Press.

Mercer, J. R., & Lewis, J. (1978). Technical Manual: System of Multicultural Pluralistic Assessment. New York: Psychological Corp.

Messick, S. (September 1979). Test validity and the ethics of assessment. Paper presented at the annual meeting of the American Psychological Association, New York.

Miele, F. (1979). Cultural bias in the WISC. Intelligence, 3 (2), 149-164.

Nerviano, V. J. (1974). The factorial structure of the Wechsler Adult Intelligence Scale: A critical reevaluation. (Doctoral Dissertation, University of Kentucky). Dissertation Abstracts International, 1975 (December), 36 (6-B) 3060. University Microfilms No. 75-26, 472, 126.

Oakland, T., & Feigenbaum, D. (1979). Multiple sources of test bias on the WISC-R and Bender-Gestalt test. Journal of Consulting and Clinical Psychology, 47 (5), 968-974.

Overall, J. E., & Levin, H. S. (1978). Correcting for cultural factors in evaluating intellectual deficit on the WAIS. Journal of Clinical Psychology, 34 (4), 910-915.

Paulsen, M. J., & Lin, T. T. (1970). Predicting WAIS IQ from Shipley-Hartford scores. Journal of Clinical Psychology, 26, 453-479.

Peteroy, E. T. (1980). Prediction of WIAS scores from Quick Test scores for white and black patients at a mental health center. Psychological Reports, 47 (1), 259-262.

Plake, B. S. (1980). A comparison of a statistical and subjective procedure to ascertain item validity: One step in the validation process. Educational and Psychological Measurement, 40 (2), 397-404.

Reschly, D. J. (1978). WISC-R factor structures among anglos, blacks, chicanos, and native american papagos. Journal of Consulting and Clinical Psychology, 46 (3), 417-422.

Reschly, D. J., & Sabors, D. L. (1979). Analysis of test bias in four groups with the regression definition. Journal of Educational Measurement, 16 (1), 1-11.

Reynolds, C. R. (1982). Methods for detecting construct and predictive bias. In R. Berk (Ed.), Handbook of methods for detecting test bias. Baltimore: Johns Hopkins University Press.

Reynolds, C. R., & Gutkin, T. B. (1980). A regression analysis of test bias on the WISC-R for anglos and chicanos referred for psychological services. Journal of Abnormal Child Psychology, 8 (2), 237-243.

Reynolds, C. R., & Hartlage, L. (1979). Comparison of WISC and WISC-R regression lines for academic prediction with black and with white referred children. Journal of Consulting and Clinical Psychology, 47 (3), 589-591.

Rudner, L. M., Getson, P. R., & Knight, D. L. (1980). A monte carlo comparison of seven biased item detection techniques. Journal of Educational Measurement, 17, 1-10.

Russell, E. W. (1972). WAIS factor analysis with brain damaged subjects using criterion measures. Journal of Consulting and Clinical Psychology, 39, 133-139.

Sandoval, J. (1979). The WISC-R and internal evidence of test bias with minority groups. Journal of Consulting and Clinical Psychology, 47 (5), 919-927.

Sandoval, J., & Miile, M. W. (1980). Accuracy of judgements of WISC-R item difficulty for minority groups. Journal of Consulting and Clinical Psychology, 48 (2), 259-263.

Sattler, J. M. (1979). Standard intelligence tests are valid instruments for measuring the intellectual potential of urban children: Comments on pitfalls in the measurement of intelligence. Journal of Psychology, 102 (1), 107-112.

Sattler, J. M., & Kuncik, T. M. (1976). Ethnicity, SES, and pattern of WISC scores as variables that effect psychologists' estimates of effective intelligence. Journal of Clinical Psychology, 32 (2), 362-366.

Saunders, D. R. (1959). On the dimensionality of the WAIS battery for two groups of normal males. Psychological Reports, 5, 529-541.

Scarr, S. (1978). From evolution to Larry P., or what should we do about IQ tests. Intelligence, 2 (4), 325-343.

Scheuneman, J. A. (1975). A new method of assessing bias in test items. Paper presented at the Annual Meeting of the American Educational Research Association, Washington (April).

Scheuneman, J. A. (1979. A method of assessing bias in test items. Journal of Educational Measurement, 16 (3), 143-152.

Schmeisser, C. B., & Fergusen, R. L. (1978). Performance of black and white students on test materials containing content based on black and white cultures. Journal of Educational Measurement, 15 15 (3), 193-200.

Semler, I., & Iscoe, I. (1966). Structure of intelligence in negro and white children. Journal of Educational Psychology, 57, 326-336.

Shepard, L., Camilli, G., & Averill, M. (1981). Comparison of procedures for detecting test item bias with both internal and external ability criterion. Journal of Educational Statistics, 6 (4), 317-375.

Smith, A. L., Hays, J. R., & Solway, K. S. (1977). Comparison of the WISC-R and Culture Fair Intelligence Test in a juvenille delinquent population. Journal of Psychology, 97 (2), 179-182.

Sowell, T. (1979). New light on IQ. In C. Torniero (Ed.) Annual Editions: Readings in Human Development 78/79. Guilford CT: Dushkin Publishing Group Inc.

Sprague, R. L., & Quay, H. C. (1966). A factor analytic study of the responses of mental Retardates on the WAIS. American Journal of Mental Deficiency, 70, 594-600.

Terman, L. M., & Merrill, M. A. (1960). Stanford-Binet Intelligence Scale: Manual for the third revision, Form L-M. Boston: Houghton Mifflin.

Vance, H. B., Hankins, N., & McGee, H. (October 1979). A preliminary study of black and white differences on the Revised Wechsler Intelligence Scale for Children. Journal of Clinical Psychology, 35 (4), 815-819.

Wechsler, D. (1955). Manual for the Wechsler Adult Intelligence Scale. New York: Psychological Corp.

Wechsler, D. (1958). The measurement and appraisal of adult intelligence (4th Edition). Baltimore: Williams and Wilkins.

Wechsler, D. (1981). Manual for the Wechsler Adult Intelligence Scale - Revised. New York: Psychological Corp.

Williams, R. L. (1970). Black pride, academic relevance and individual achievement. Counseling Psychologist, 1 (1), 18-22.

Williams, R. L. (1971). Abuses and misuses in testing black children. Counseling Psychologist, 2 (3), 62-73.

Williams, R. L. (1974). The problem of the match and mismatch in testing black children: A symposium. Englewood Cliffs, N.J.: Prentice Hall.

Williams, R. L. (1975). The BITCH-100: A culture specific test. Journal of Afro-American Issues, 3, 103-116.

Wright, B. D., & Stone, M. H. (1979). Best test design. Chicago: Mesa Press.

Zaref, L., & Williams, P. (1980). A look at content bias in IQ tests. Journal of Educational Measurement, 17, 313-322.