8-1982

An Investigation of a Procedure to Assess Aggregated Item Bias in a Minimum-Competency Test

Laurence E. Rudolph
Western Michigan University

An Investigation of a Procedure to Assess Aggregated Item Bias in a Minimum-Competency Test

by

Laurence E. Rudolph

A Dissertation
Submitted to the
Faculty of the Graduate College
in partial fulfillment of the
requirements for the
Degree of Doctor of Education
Department of Educational Leadership

Western Michigan University
Kalamazoo, Michigan
August 1982

AN INVESTIGATION OF A PROCEDURE TO ASSESS AGGREGATED ITEM BIAS IN A MINIMUM-COMPETENCY TEST

Laurence E. Rudolph, Ed.D.

Western Michigan University, 1982

This study described and evaluated a new procedure to assess item bias in a minimum-competency test (MCT). This procedure was thought to be capable of estimating the degree of bias contained in a given test item. This is in contrast to traditional item bias detection procedures which focus on the presence or absence of bias in an item. Furthermore, the procedure was thought to be capable of estimating item bias in such a way that the aggregate of item bias (AIB) could be obtained.

The chi-square was found to be the most practical of the item bias procedures and the AIB uses a member of this family. The AIB computes a phi correlation coefficient, squares this to obtain an estimate of the coefficient of determination which estimates the percentage of item

variance which is attributable to the demographic characteristic used to separate the subgroups. The coefficient of determination is multiplied by the variance of the item to obtain an estimate of bias in the item and each resulting number is aggregated across test items to obtain an estimate of total item bias in the test.

The AIB was compared to an adaptation of a traditional item bias detection procedure (SSTD) computed on the same data set. The SSTD was a simple directional count of biased items determined by the chi-square procedure. Seven estimates were obtained for both procedures on subgroups which were randomly constructed. An F-test was made on the ratio of the variances of the two procedures. The AIB was found to be the more stable of the two procedures. Subsequent analyses showed that the AIB and SSTD were consistent in their determination of the direction of bias (against Black students) and this was in agreement with population data showing Black students scoring one standard deviation below White students. A t-test of results within the two procedures across the two score points showed that it was allowable to combine these results in the analyses, i.e., scores one and two points below the score which defines a minimally competent student. The AIB was found to be a promising alternative to traditional bias detection procedures.

Rudolph, Laurence Eugene

AN INVESTIGATION OF A PROCEDURE TO ASSESS AGGREGATED ITEM BIAS IN A MINIMUM-COMPETENCY TEST

Western Michigan University ED.D. 1982

ACKNOWLEDGEMENTS

The Dallas Independent School District (DISD) graciously allowed me to access their superb data base to obtain the information which was basic to this research effort. They further donated the computer time for the bulk of the data analysis for this study. I certainly could not have completed this study without their assistance. I would like to specifically thank Drs. LaVor Lym and William Webster of the DISD for the kindness they showed to me during the time I spent there.

Support was also given by the Evaluation Center of Western Michigan University. The Center provided office space, computer facilities and most of all friendship throughout the process to complete this research. The opportunity to work at the Center was one of the best things that has happened to me. I learned much during the years I spent there.

I was very lucky to have chosen Western Michigan University for my graduate studies. The College of Education, Graduate College and Department of Educational Leadership were always available and willing to assist. The library and computer facilities were all one could ask for. Thankyou, WMU.

This document reflects the effort and support of a number of special people. David Vines wrote the computer programs to do the data analysis for this study. He is a singular individual. His skill with the computer made an arduous process easy. Thanks, Dave. Reagan Sides helped in the editing of this document, he kept me sane (somewhat). Jeri Rydings kept telling me I could do it.

I received guidance and support from the members of my committee, Dr. Galen Alessi and Dr. James Sanders. Any value derived from this study is to a large measure due to the input of these men. They showed a balance of tolerance and professionalism that greatly aided the process and product.

Dr. Mary Anne Bunda, the chair of my committee, was truly a joy to work with. I am not sure that I could have completed this document without her friendship and

support. There is a warm spot in my heart for her. I hope that I can make her proud of me.

Wendy Leys was my pillar of strength throughout this process. Everyone told me that the writing of a dissertation would threaten our relationship. Her support and assistance brought us closer together and made me love her even more.

Chloe sat at my feet and wagged her tail and made me feel a little better as I sat at the computer terminal at three in the morning. I owe her many walks in the woods and a bushel of Liva Snaps.

In the back of my mind I know that one of the reasons for my completing this document was to please my family. Dave and Millie are two of the most wonderful people in the world. I sit here and think of them and feel that there is reason to hope for the future and that I can be part of that which they have worked for. Steve, Maria Elena, Nicky and Scott are a wonderful foursome to model. I love you all.

Laurence Rudolph

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

STATEMENT OF THE PROBLEM

The use of criterion-referenced tests (CRTs) is gaining in popularity in American schools. CRTs are designed to assist in the making of dichotomous decisions about test takers: competent or incompetent, master or nonmaster, pass or fail, and so on. One of the uses of a CRT may be as a minimum-competency test (MCT); for example, a score on an MCT may be used as a requirement for graduation from a school system. Students who score below a specified level are considered "incompetent" and denied graduation. Students at or above the score are considered "competent" and allowed to graduate. This sort of decision is of such importance that extreme care must be taken to ensure that the test used is valid and reliable for that purpose.

Because CRTs and MCTs are used to make absolute rather than relative decisions on performance, the indices for reliability and validity should differ from traditional (norm-referenced) procedures. Rather than measuring test score stability, reliability indices for CRTs and MCTs focus on the stability of decisions made with the test (Swaminathan, Hambleton & Algina, 1974). Test score instabilities that do not effect decisions are not seen as a significant factor in the reliability of a CRT. (Note: the use of the word "effect" is deliberate and reflects the absolute decisions based on test scores.)

Validation of CRTs requires the consideration of a number of factors pertaining to the construction, administration and use of the tests. Among these is that the test must discriminate between students based upon the skills that the test is intended to measure and not discriminate between students based upon irrelevant factors such as demographic characteristics (e.g., race or gender). Any test which discriminates along one of these irrelevant dimensions is considered to be biased with respect to that dimension. Bias, then, is seen as a threat to validity in a test when it is used to measure the skills of students who vary along a dimension upon

which the test has been shown to be biased. In other words, interpretations of test scores in a biased instrument are ambiguous due to lack of specific knowledge of the interaction of competency and bias on test score. Such ambiguity is unacceptable in a test used to make decisions as important as graduation from high school.

Bias in tests has traditionally been investigated in two areas: test bias and item bias. Test bias has been defined as the measured differences among test scores of demographically diverse students of equal competency and is attributable to individual differences along the irrelevant demographic characteristic (Shephard, Camilli & Averill, 1980). Item bias has been defined as the measured difference on a specific item between equally competent students who vary demographically (Shephard et al., 1980). This study will focus on item bias and will do so with a perspective similar to that of Swaminathan et al. (1974) on reliability, in that bias which does not effect decision making will not be included in the estimated bias of an item on an MCT. This focus will later serve as a basis for the selection of a particular level of competence in the sample studied.

Item bias procedures have been useful in decision making relative to an individual item on a test. Procedures have been developed to identify specific items which are likely to contain item bias (e.g., Angoff & Ford, 1973; Green & Draper, 1972; Ironson & Subkoviak, 1979; Lord, 1980; Scheuneman, 1979; and Wright, 1977). These procedures can be used to determine which items on a test should be considered for alteration or exclusion. None of the procedures, however, is useful in determining the impact of item bias on the total test score or on decisions based upon that score. They are intended only as a means of flagging items which are likely to contain bias. None allows for the aggregation of bias across items to estimate the total item bias in the test or as a means of comparing tests on the basis of total item bias. Consequently, if conclusions concerning test bias are to be made on the basis of item information, procedures must be developed and tested.

Objective of the Study

The objective of this study is to present and evaluate a new procedure for the assessment of bias in items on minimum-competency tests. The presented

procedure is similar to other item bias procedures in that it uses a statistic from the chi-square family of non-parametric statistics. The procedure is different in that rather than attempting to determine whether or not bias is present in a given item, this procedure will attempt to estimate the degree of bias in the item. Furthermore, the bias estimates are thought to be capable of being aggregated across test items on a minimum-competency test in such a way that inferences can be made as to the aggregated item bias contained in the test.

The evaluation of the presented bias assessment procedure will compare the stability of obtained results to those obtained with a traditional bias procedure. Stability is a relative term. For a bias procedure to be of use, it must be stable. If the presented bias procedure is found to be more stable across applications than traditional bias detection techniques, it might be a more useful approach to the detection of biased items. Also, since the bias in the items is measured in such a way that it can be aggregated across all the items on the test, further use could be made of the procedure.

Traditional item bias detection procedures inspect some aspect of performance on an item and make a determination as to whether the item is biased. No inference can be made as to the degree of bias in the item. The only way to aggregate such information across items on a test is to count the number of biased items that were found on the test. Comparisons of tests based on amount of bias may not be directly determinable from the number of biased items on the test. This is because items may contain different degrees of bias, and this bias, when aggregated, may have greater or lesser total impact on the total test score.

In light of these factors, the objective of this study is to present and evaluate a new item bias detection procedure, one that generates bias estimates which can be aggregated across items in the test. A parallel application will be made of a traditional item bias detection procedure. The two procedures will be applied to a data set and their stabilities compared.

CHAPTER II

REVIEW OF THE LITERATURE

The objectives of this review of the literature are: to set the study in an historical perspective; to establish a framework for the study; to define terms used in the study; and to establish a conceptual background upon which the study can be evaluated.

Types of Tests

This review will cover three types of tests, norm-referenced, criterion-referenced and minimum-competency. Each of these refer more specifically to the way in which a score is used rather than some characteristic of the test itself. Examination of a test item would not yield information concerning the way in which the score from that item was used (Popham & Husek, 1969), yet the use of the score determines to a great degree the technical quality of the item.

Norm-Referenced Tests

Norm-referenced tests (NRTs) are tests developed to make relative judgements about test-taker performance in relation to some comparison group (Hambleton, Mills, Simon & Livingston, 1980). Because of this, items on an NRT are chosen (partially) on the basis of their ability to differentiate student performance. An item which does not produce variability in student performance will most likely be rejected (Berk, 1981). Relative decisions cannot be made based upon item performance with no difference between students. The item would be rejected irrespective of the educational significance of its content.

Norm-referenced tests can be designed to offer information specific to different comparison groups. A score from a test can be viewed with regard to all test takers and/or some subset of test takers (e.g., males, ten-year-olds or private school students). This information can be used to make inferences about different individual students and subgroups of students relative to other students and subgroups of students. Much of the testing which done in the public schools is of this type. A school district can compare the performance of the their students

to other school districts with similar demographic characteristics. Additionally, because of the variability in the test scores, these tests are easily used in correlational or predictive studies of later behavior.

Criterion-Referenced and Minimum-Competency Tests

Criterion-referenced tests (CRTs) and minimum-competency tests (MCTs) will be discussed together. This is due to the similarity of these two testing strategies. Minimum-competency tests will be viewed as a subset of CRTs. At times literature from CRTs will be used to describe MCTs. In such cases the language will speak of CRTs to reflect the content of the source; however, the reader should be aware that the information is intended to support the understanding of MCTs.

Glaser (1963) was the first to mention a need for an alternative to norm-referenced testing. Norm-referenced tests (NRTs) were seen as limiting when making decisions on educational programming for students.

> Norm-referenced tests are constructed, principally, to facilitate the comparison of individuals (or groups) with one another or with respect to a norm group in the content area measured by the test. Criterion-referenced tests...are constructed to permit the

> interpretation of individual (and group) test scores in relation to a set of clearly defined objectives and competencies (Hambleton et al., 1980, p. 3).

What has evolved over the eighteen years since Glaser's paper is a separate body of test development procedures whose objectives are considerably different from those of norm-referenced test development procedures. The objective of CRTs is to compare an individual's performance to some criterion: "For example, the dog owner who wants to keep his dog in the back yard...wants to find out how high the dog can jump so that he can build a fence high enough.... How the dog compares to other dogs is irrelevant" (Popham & Husek, 1969, p. 2). Since NRTs are intended to accertain an individual's performance relative to others, applications of NRTs will differ from those of CRTs (Popham & Husek, 1969). Because of these different objectives, items chosen for inclusion on CRTs and MCTs do not consider variability of student performance. If the content of an item is thought to be important, it will likely be included in the test even if all students answer it correctly.

Citing a definition for criterion-referenced testing is not a clear task. Gray (1978) listed fifty-seven definitions for criterion-referenced testing obtained from the literature. Although the differences between these definitions are often minor, one distinction seems important. Some definers of criterion-referenced testing state that the word "criterion" in CRTs refers to the use of the test to make decisions (based on a criterion or cutoff score) with regard to the competence of the test taker (see Sanders & Murray, 1976). Other definers (e.g., Millman, 1974) focus on the criterion as a domain of content or behavior to which test scores can be referenced. Although it is the second use of criterion (domain of content) that has won favor among many professionals in the field of criterion-referenced testing, it is the first use of criterion (specific score) that is most relevant to this study. This domain of content differs from the use of domain in domain-referenced testing, in that domain-referenced testing results in a probabilistic estimate of the inferential power of the sample to the domain. The use of probabilistic estimates in most criterion-referenced tests is non-existent.

Since this study deals with an MCT, the view of a criterion as a cut-score is applicable. What is important to note is that a test is criterion-referenced due to its construction (referenced to a domain of content) and a test is an MCT due to its application (used in an absolute decision-making process). A test which was developed to be norm-referenced could be used as an MCT, although this practice might not be condoned by experts. Norm-referenced tests are used to make relative decisions about test takers, such as their performance compared to other students at a given grade level. Criterion-referenced test are used to make decisions based upon some set of skills -- the student is capable of accurately computing eighty percent of a set of division problems. Minimum-competency tests are used to make absolute decisions for students -- students with scores above 45 will be allowed to participate in advanced classes. Consequently, it seems unlikely that a school district would set, as a criterion, a score below which fifty percent of the students scored and yet not know the exact characteristics of the skill measured. For more on criterion-referenced testing and minimum-competency testing, see Hambleton et al. (1980) and Bunda and Sanders (1980).

The use of MCTs is becoming widespread in the United States. As of 1980, thirty-one of the fifty states had passed legislation involving the testing of students before graduation (Berk, 1980). This has been a continuing trend over the past decade. This increase was attributed to public pressures on education for accountability in student achievement, specifically in the granting of high school diplomas (Berk, 1980).

Introduction to Bias in Testing

Much of the current and prior work in item bias detection has focused on intelligence testing. A smaller, but still significant, body of work has focused on item bias in norm-referenced achievement testing (e.g., the Scholastic Aptitude Test). It has been found that the construction and content of an item can greatly impact the bias of an item. Items containing negatives in the stem (e.g., "Which of the following is not a vehicle?") have been shown to be biased against Black students (Linn & Harnish, 1979). In the 1950's an item which used a modern desk phone as an illustration was shown to be biased against rural students (Jensen, 1980). In the first

instance, an item which was intended to measure student understanding of the concept "vehicle" appeared also to measure the students' familiarity with the use of negatives. In the second instance, lack of familiarity with the type of telephone used in the illustration biased the item. In both instances students who may have had the skill that the item was intended to measure may have incorrectly answered the item due to irrelevant factors. This is clearly a form of bias in a test item.

All biased items are not able to be detected by the inspection of the item. Subtleties in the item composition may bias an item yet elude inspection for biased content. In such cases it is only after students have taken the test and their performance is inspected in some context that an item can be seen to be biased. The procedures developed to detect this bias form a framework for the current study. No study of item bias specific to CRTs or MCTs was found in the literature, nor was any mention made of differences in item bias detection procedures for CRTs or MCTs. Investigation into this area might help to expand knowledge of possible differences in item bias detection in these tests. Differences are seen between other procedures. A number of papers in the

literature suggest that different techniques should be applied to CRTs for reliability (Linn, 1980; Millman, 1974; Hambleton & Novick, 1973; and Swaminathan et al., 1974). An index of the reliability specific to CRTs has been developed; it is the kappa coefficient (Swaminathan et al., 1974). The kappa indicates the consistency of decision in the use of a test, not the consistency of score obtained as is done with norm-referenced reliability procedures.

If CRTs have alternative indices of reliability, it may be prudent to suggest alternative techniques for the assessment of bias in CRTs and MCTs. This study seeks to apply the perspective of consistency of decision (Millman, 1974) made to the study of bias in an MCT; bias that does not effect decision making is not to be included in the bias estimate for an MCT item. This study will concentrate on the methodologies of bias detection, not on the substance of item content.

Definition of Bias

Bias in testing is a complex concept to define. In part, the definition of bias will depend upon or affect the procedure used to assess bias (Shephard et al., 1980). Generally, bias is invalidity that differentially impacts one subgroup. The presence of bias, as with all types of validity, is dependent upon the way in which a test is used (Cronbach, 1971). Therefore, bias must be viewed within a specific context. A test may be biased in one application and unbiased in another application.

Bias in testing has been broadly divided into two types: test bias and item bias. The definitions as cited by Shephard et al. (1980) are relevant to the perspectives of this study:

> A test is biased if equally able individuals, from different groups, do not have equal probabilities of success...an item is biased if two individuals with equal ability but from different groups do not have the same probability of success on the item. (pp. 2-3)

The authors continue:

> This understanding [definition of item bias] is difficult to operationalize because simple inspection of group differences leaves us uncertain as to which are true differences

> and which are evidence of bias. The process and reasoning followed in detecting bias is much like that used in construct validation. It is the pattern and entire network of relationships that support or disconfirm validity. Similarly, bias cannot be identified in an isolated test item. Test questions designed to measure the same construct must be studied together; bias is discovered when an item does not fit the pattern established by others in the set. Bias is a contextual property; it depends on the characteristics of the items comprising the test that are used to measure the "ability" in the definition of item bias. Thus, the "bias" assessed by these techniques is "anomaly in a context of other items".... (pp. 3-4)

This study will take a similar view of item bias, and when aggregated across items on the test, to test bias. By making certain assumptions, anomalies in item performance can be directly interpreted as bias, and assigned a value which can be aggregated to derive an estimate of bias in the test. However, unlike Shephard et al., within this study performance on a given item will be viewed independently from other item performance. The impact of item intercorrelations is excluded in this perspective, but the capability of estimating the degree of bias in an item may offset this exclusion.

Item Bias Techniques

More than an dozen techniques have been developed to detect bias in test items (Shephard et al., 1980). Of these techniques, six have received support in the literature:

1. Three-parameter item characteristic curve.
2. One-parameter item characteristic curve.
3. Transformed difficulty.
4. Item discrimination indices.
5. Chi-square using correct responses only.
6. Chi-square using correct and incorrect responses.

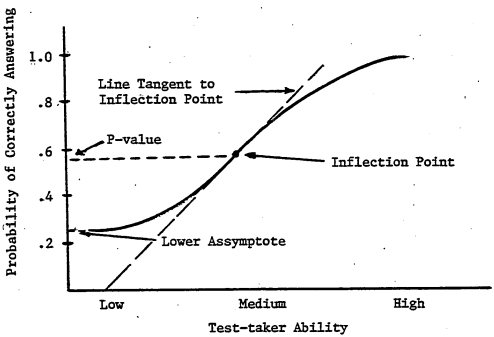
Each of these item bias detection procedures will be described below.

Three-Parameter Item Characteristic Curve (ICC-3)

The ICC-3 is the most sophisticated of the item bias detection procedures (Lord, 1977). It is also the most complex and costly both in computational time and required sample size (the ICC-3 requires 1,000 subjects and a test of at least 40 items). The theory in which performance and bias are described is called latent trait theory. The

ICC-3 gives a graphic display of the relationship of test taker ability and the probability of the test taker correctly answering the item. The three parameters that define the curve are the discriminating power of the test item, the overall difficulty of the item, and the probability of a student with a low skill level correctly answering the question by chance. Figure 1 is a sample of what a ICC-3 curve and tangent line might look like.

Figure 1.
ICC-3 Illustration



The discriminating power of the item is defined as the slope of the line tangent to the curve at the point of inflection. A steeply sloped tangent line indicates high discrimination power -- a small difference in test-taker ability greatly changes the probability of correctly answering the question. A tangent line with a less steep slope indicates lower discriminating power; test takers with a wide range of ability have approximately the same probability of getting the item correct.

The difficulty of the item is expressed as the point of inflection of the ICC-3 curve. The point of inflection is the point on the ICC-3 curve where the second derivative of the curve changes from positive to negative. The higher the point of inflection is in the curve, the easier the item. The lower the point of inflection is, the more difficult the item.

The probability of a test taker with low skills correctly answering the test item is displayed in the height of the curve at the lower asymptote. If the lower asymptote of the curve is at .25, a student with low skills would have a one-in-four chance of correctly answering the question (such as might be the case in a

four-option multiple choice test).

Indices have been developed using the ICC-3 to investigate differences between any of these parameters, also with parameters grouped in different ways. One of the more popular methods has been to determine ICC-3 curves for two subpopulations and to measure the area between the two curves. The area between the two curves is said to express the bias in the item (Lord, 1977). One difficulty with this method is that it does not produce an estimate of the probability that the area measured might be due to chance differences between the curves. This lack of an informed process for determining whether differences in performance on an item are due to real biases or due to chance differences in performance limit inferences which can be made as to the true bias of the item. An index developed by Wingersky (1977) combined the slope (discrimination) and inflection point (difficulty) of the ICC-3 curve. This procedure assumes that chance levels for the two subpopulations are equal. This is an ICC-2 method. The result is an index which can be tested with the chi-square, a non-parametric statistic with a significance estimate.

One-Parameter Item Characteristic Curve (ICC-1)

The ICC-1 (Duvoric, 1975) is a simplification of the ICC-3. It is conceptually similar to the ICC-3, but is less complicated due to the use of only one of the parameters used in the ICC-3 method. The ICC-1 method assumes that there is no guessing on the test (all correct answers are due to test-taker knowledge of the item) and that all items are equally discriminating. This approach is somewhat similar to the simplifying assumptions made to develop the KR-21 reliability index by assuming equally difficult items in the KR-20 index.

The most often used index for the ICC-1 is the difference in probability estimates for the two subgroups. Wright, Mead and Draba (1976) developed a way to test the fit (accuracy) of the curve to the model. The measured probability of a person of a given ability (derived from the ICC-1 curve) is subtracted from that person's score on the item. For example, a person of a given ability (based on some criterion external to the item performance) might be estimated to have a 75% chance of answering a particular item correctly. If that person correctly answers the item, he/she is given a score of 1.0 on the

item. The probability of that person correctly answering the item (.75) is subtracted from the score (1.0) on the item. The result (.25) is called a residual. For a given subgroup on a given item, all of the residuals are squared and then summed. This sum of the squared residuals is then divided by the number of persons in the subgroup to obtain an estimate of the mean squared residuals. The same procedure is performed for another subgroup. The estimates for the mean squared residuals are compared. The presence of bias in the item is estimated from the observed difference between the values for the mean square residuals for the two subgroups. If a relatively high value is found for the mean square residual, the item is assumed to be biased. The are no objective measures for determining the significance of the observed difference in the mean square residuals. This is seen as a limitation of this item bias procedure.

The usefulness of the ICC-1 is limited by the validity of the simplifying assumptions. If the items on a test are equally discriminating and guessing factors are zero, then the ICC-1 will be equivalent to the ICC-3 with considerably less calculation. These assumptions are, however, difficult to meet in the field.

Chi-Square Using Correct Scores Only (Chi-Square+)

The chi-square+ procedure was developed by Scheuneman (1979) to measure item bias in a fashion similar to ICC-3 but in a less complex and demanding way. The chi-square is a non-parametric procedure that assesses bias by considering student performances with the expectation that students with the same or similar total test score should do equally well on a given test item irrespective of the demographic characteristics of the student.

In the chi-square+ procedure students are divided into balanced subgroups based upon total test score and a demographic characteristic, for example, race. These subgroupings usually result in five bands of students by test scores. The assumption underlying the chi-square+ is that students within each of the bands of test scores are of equivalent knowledge in the content area measured by the test and should perform equally well on each item on the test. A set of expected frequencies is determined for each band of students. This set of expectations is compared with the actual performances of the students. Any difference between the observed performance and the expected performance is squared and divided by the

expected performance. The quantities are summed across all of the bands of the students. The sum of these quantities is distributed as a chi-square. The sum is compared to the distribution of chi-squares with the appropriate number of degrees of freedom and an estimate is obtained of the probability that the difference in student performance is due to chance. The degrees of freedom are calculated by multiplying the number of subgroups minus one by the number of score bands minus one.

The chi-square+ was found to be flawed in the way it responded to items with different difficulty levels (Camilli, 1979). The difficulty levels are referred to as p-values. These values are used in classical test theory to describe the proportion of test takers who correctly answer a given question. This differential performance of the chi-square+ was said to be caused by the lack of inclusion of incorrect student performance and an unequal balance of demographic subgroups.

Chi-Square Using Correct and Incorrect Responses (Chi-Square+,-)

Camilli (1979) devised a procedure where both correct and incorrect student responses are included in the chi-square calculation. For each of the bands used in Scheuneman's chi-square+, the chi-square+,- computes an individual value using both correct and incorrect scores. This is repeated over all of the bands previously included in the chi-square+. A value is computed for each of these bands of students. The sum of these values is compared with the same chi-square significance tables as the chi-square+ procedure. The advantage of this system over the chi-square+ is that it is not sensitive to differences in p-values. The disadvantage of this method is that it requires the computation of more than one chi-square (most likely five) for each item, although only one probability estimate is still found.

Item Discrimination Index (IDI)

The IDI uses the point-biserial correlation measured between an item and the total test score for different subgroups. Point-biserial correlations are often used in

test development to screen items. The point-biserial is the appropriate correlation between the performance of a student on an item with the student's total test scores. The idea is that any item that does not correlate with the total score is not a good item or is measuring a skill different from the skills measured by the test. This is the traditional index for item discrimination used in work which has its base in classical test theory rather than latent trait theory.

Classical test theory is a set of propositions originally developed by Thurstone which state that any observed score on a test is an indicator of the test taker's "true" standing on the ability which the test measures. However, observed scores carry with them a component of measurement error which is random. Under this rubric, test items are viewed as indicators of the ability of the individuals who have taken the test. Thus characteristics of items are designated statistics rather than parameters and are likely to change when the test is given to other groups of students.

Latent trait theory views item characteristics as a set of parameters which indicate test-taker standing on a given trait and will not change with a different groups of students. Thus, the point-biserial correlations in the IDI are interpreted similarly to the slope of the tangent line in the ICC-3. They are different in that the point-biserial correlation is expected to change across groups of students and the slope of the ICC-3 is not.

In applying the IDI to the measure of item bias, differences in point-biserials for two subgroups are compared. An item should have the same or similar point-biserial for different subgroups. Green and Draper (1972) stated that an item was biased if it were in the top half of the test items when ranked by IDI for one subgroup and the lower half for another subgroup. The degree of difference within the distribution is said to reflect the amount of bias in the item.

A major difficulty was noted in the IDI. Indications of bias might be generated by mean differences between the subgroups. Since the IDI is based on the point biserial correlation coefficient, and all correlation coefficients are sensetive to variance, two subgroups with large

p-value differences on an item might indicate bias due to differences in variance within the two subgroups (Shephard et al., 1980). Variance in group performance on an item decreases as the p-value for that subgroup on that item deviates from .50. In the perspective of latent trait theory, two subgroups with large mean differences might produce high point-biserial differences even if both subgroups had identical item characteristic curves, since the point-biserial is a statistic and not a parameter. The difference between correlation coefficients is due soley to the amount of variance available in the item for each subgroup.

Transformed Item Difficulty (TID)

The TID was developed by Angoff and Ford (1973). An item was considered biased if it is relatively more difficult for one subgroup than for others. The p-values for each item for each subgroup were transformed to standard scores (z's) based upon the standard deviation of the p-values on the entire test for each of the subgroups. Differences in z-values between the subgroups on a given item indicate the significance of bias. Because the TID does not take into account differences in subgroup means

for the total test, bias may be accentuated or masked due to real differences between the two subgroups (Lord, 1977). The TID is therefore most useful in studies where the subpopulations have similar group means on the test.

Comparisons of Item Bias Procedures

Four studies comparing item bias procedures were reviewed. One study, Rudner, Getson and Knight (1980), used simulations to generate biased items. The other three studies (Ironson & Subkoviak, 1979; Rudner & Convey, 1978; and Shephard et al., 1980) used actual student responses. Rudner et al. (1980) used a computer to simulate biased items in a data set. This simulation was based upon the ICC-3. Some of the items on a data set were constructed to reflect bias. Bias detection procedures were evaluated based upon agreement with the simulation. The percentage of agreement on biased and non-biased items was compared. The other studies were similar in their comparisons but they used data from actual student performance. Rather than comparing the outcome of the different procedures to known bias (bias that was constructed), these studies compared the agreement of the procedures to the outcome of the ICC-3

calculations. The ICC-3 was the standard for comparison.

In general the findings of all four studies indicated the ICC-3 to be the method of choice. However, since the ICC-3 was the method used to generate the biased item in the study by Rudner et al. (1980), it was not surprising that it was found to be the best method. What was interesting, though, was that the ICC-3 was able to detect only 80% of the biased items. In the other three studies, those that used actual student test performance, the ICC-3 was the model used to check on the accuracy of the other methods. It is considered the most elegant of the bias detection methods, actually superior to item bias techniques which apply classical test theory statistics (Shephard et al., 1980).

All four of the studies found the chi-square procedures to be the best alternative to the ICC-3.

> The three-parameter methods [ICC-3] are theoretically the most sound but cannot be used properly without samples of at least 1,000. Given that bias detection techniques are never to be used as hard and fast criteria for rejecting test items, it is safe to recommend that the Scheuneman chi-square [chi-square+] with the Camilli modification [chi-square+,-] could be used as the best substitute for the ICC-3.... (Shephard et al., 1980, p. 75)

Two techniques -- item characteristic curve

> with three parameters and chi-square with five intervals [Scheuneman, 1979] -- produced fairly accurate results under all investigated conditions. (Rudner et al., 1980 p. 9)

Each of the other techniques was found to be lacking, either in its sensitivity to irrelevant factors or its insensitivity to relevant factors.

Summary

There are distinct differences between norm-referenced tests and criterion-referenced and minimum-competency tests. Because of these differences alternative indices of reliability are used. This is seen as support for the development of a different index for the estimation of bias in MCTs. If this index is intended to parallel the differences between reliability in NRTs and MCTs, it should focus on consistency of decision not on consistency of test score.

A number of different item bias detection procedures are available. The one most supported by the literature is the ICC-3 (Lord, 1977). However, the ICC-3 is very costly in computer time and requires a large sample size to obtain stable estimates. The best alternative to the

ICC-3 is the chi-square+,- (Camilli, 1981).

CHAPTER III

METHODOLOGY

This chapter details the logic behind the presented bias detection procedure, and describes that procedure and a study to investigate its stability.

Development of an Aggregable Bias Technique

The stated objective of the study is to present and evaluate a procedure to estimate item bias that can be aggregated across test items to obtain an estimate of the aggregated item bias (AIB). To accomplish this, a shift in focus from traditional item bias detection is necessary. No attempt is made in traditional item bias detection procedures to combine the results across items on a test to estimate the impact of item bias aggregated across all items in the test. This is due in part to the

focus of these procedures on the determination of the presence of bias through a probability estimate. This probability estimate is not directly interpretable as an estimate of the amount of bias in the item, but rather a statement of the certainty that the item is indeed biased. The item bias procedures of classical test theory, particularly the chi-square, are affected by the number of observations included in the calculations. The larger the number of observations included, the more sensitive the procedures are to small differences in the dependent measure (i.e., student performance on the item). Computed probability estimates can be low (significant) due to the strength of the relationship or due to the number of observations included in the calculations. Therefore, no attempt is made to infer the degree of the bias through these probability estimates.

If an estimate of the degree of bias in an item is desired, what is needed is a statistic that directly investigates the strength of the relationship between demographic characteristics (irrelevant dimensions) and item performance. This procedure would not consider bias as an all-or-none characteristic of an item but as a characteristic of an item which can vary by degree. Such

a statistic exists: it is the phi statistic and is derived from the chi-square, an often-used procedure for the detection of item bias (e.g., Scheuneman, 1979). Phi adjusts for the number of observations included in the chi-square and results in an estimate of the strength of the relationship between the independent variable (demographic characteristic) and dependent variable (test item performance). One procedure for computing phi is to take the square root of the computed chi-square and dividing it by the number of observations included in the computation of the chi-square.

Phi takes on the value of zero when no relationship exists between the variables (demographic characteristic and item performance) and departs from a value of zero when the variables are related. "The meaning of the phi coefficient in a correlational sense is quite clear; it is simply the Pearson product-moment coefficient of correlation for dichotomous data" (Glass & Stanley, 1970, p. 161). By squaring the computed phi coefficient, the coefficient of determination is obtained. The coefficient of determination is "the proportion of variance in the dependent variable that is predictable from the independent variable" (Hopkins & Glass, 1978, p. 160).

This is precisely the type of measure that is necessary for the line of reasoning pursued in this study.

Phi is not without its limitations and problems. Shephard et al. (1980) list two difficulties with the use of the chi-square (the family of non-parametric statistics of which phi is a member) in the assessment of item bias that affect the present study:

> 1. Due to differences in group means within the matched intervals, regression artifacts can still cause the appearance of bias (p. 120).
>
> 2. ...the procedure is confounded by unequal sizes for the two groups (p. 120).

A third difficulty pertains specifically to the calculations of phi:

> 3. When the distribution of the scores within the phi calculation are extremely disproportionate, the range of phi is restricted -- not +1 to -1 (Cureton, 1959).

The first two difficulties refer to previous applications of the chi-square to item bias (Scheuneman, 1979). The difficulty noted with regression artifacts can be avoided by computing the phi on groups of students with the same test score. The difficulty with sensitivity to unequal sample sizes can be avoided by balancing the

sample sizes, both of which will be discussed in the Subject Section below. The third difficulty, however, is inherent in the statistic. One attempt to adjust for this lack of +1 to -1 range was suggested by Cureton (1959). This adjustment involved the computation of the maximum value for phi (phi max) given the distribution of frequencies in the cells. The computed phi is divided by the value of phi max to obtain an adjusted phi which does have a range of +1 to -1. This procedure was criticized by Carroll (1961) when he demonstrated that the adjustment was unstable over the range of phi. There appears to be no acceptable procedure to avoid the difficulty of the restricted range of phi.

When the range of phi is restricted, the estimate of correlation between the independent and dependent variables will be underestimated. However, there is no reason to assume that this restriction will differentially prejudice the bias estimate towards one of the demographic subgroups. Consequently, the expectation is that a bias procedure which used the phi would (at worst) be an underestimate of bias. It would, however, reflect the true direction of the bias, if the sign of the phi were controlled. Although this is seen as a limitation to the

proposed procedure, the limitation is one of interpretability and not of substance.

Development of Procedure

Assumption of Equal Competence

Applying the phi coefficient to the present research question will involve two assumptions. The first assumption is that observed differences in performance of students separated by demographic characteristics are due to the differences in the characteristic (bias) and not due to real differences in competency. This assumption is supportable only if the students are equally competent in the skills measured by the test. Any item that has different p-values for equally competent students who differ by some demographic characteristic, is, by definition (Shephard et al., 1981), biased with respect to that demographic characteristic. Matching students by competence is a difficult task; however, in the case of MCTs where, by design of the test users, one score point separates competent from incompetent students, stringent demands can be placed upon the validity of the test, especially in the range of scores close to the cut-score.

Since decisions are made directly from test scores, students who score one score point below the cut-score can be assumed to be just short of acceptable competence. Any test not capable of making such fine discriminations of competence should not be used for that purpose. A test found to be in use for such a purpose can, by the fact of its use, be presumed capable of making fine discriminations of competence. Therefore, students achieving the same test score, if that score is close to the cut-score, can be considered equally competent. Any difference in p-values on an item, between these equally competent students, when separated by a demographic characteristic, can be attributed to bias with respect to that demographic characteristic. That is, the expected p-values for those two subgroups are assumed equal.

Given this assumption, a procedure to estimate the degree of bias in an item can be described. A phi is computed on the performance of two subgroups, and the resultant phi is squared and interpreted as the proportion of item variance which is attributable to bias with respect to the characteristic that was used to separate the two subgroups. This is the coefficient of determination, representing the degree of population

variance which is attributable to the independent variable (subgroup characteristic). By multiplying the coefficient of determination by the variance of the item performance, a value is obtained which reflects the degree of impact that the separation of the subgroups had on item performance. If these subgroups are assumed to be equally competent, this impact must reflect the bias in the item with respect to the characteristic used to separate the two subgroups. The variance of item performance is the proportion of students correctly answering the item (p) multiplied by the proportion of students incorrectly answering that item (q). The computational formula for this bias detection procedure would be:

$$IB_j = PHI^2_j \times (p_j \times q_j)$$

Where "IB sub j" is the estimated bias for the jth item on a test; "phi squared sub j" is the coefficient of determination for the jth item for that separation of subgroups; "p sub j" is the proportion of all students in the study group who answered the item correctly; and "q sub j" is the proportion of students who responded incorrectly to the item. This is the first step in the process, the presentation of a bias detection procedure

that estimates the degree of bias in an item. The next step is to define a procedure whereby the values for the item biases can be aggregated to derive an estimate for the aggregate of item bias. To do so involves another assumption.

Assumption of Direction of Bias

The second assumption deals with the determination of how the bias impacts the performance of the population. An item may be biased against the lower-performing subgroup (i.e., the difference in p-values for the two subgroups is due to the lowering of the performance of one of the subgroup); or the item may be biased in favor of the higher-performing subgroup (i.e., the difference in p-values is due to the raising of performance of one of the subgroups); or the two may be combined. There is no way of knowing (within the perspective of performance on a single item) which of these is the case. However, as stated above, this study is seeking to estimate that bias which effects decision making. For students grouped at a score of one below the cut-score, bias in favor of one subgroup (positive bias) would cause some individuals in the subgroup to get the item correct who would not have done so if the bias was not present. Adjusting for this

bias would lower the total score of some members of that subgroup. Since their scores are already below the cut-score, this adjustment would have no impact on decisions made with the test. It is only the type of bias which causes students to incorrectly answer a question they would have correctly answered had the bias not been present in the item (negative bias) that is of interest to this study. Consequently any bias in an item will be signed in such a way that it reflects a negative impact on student performance. If a given item is found to be biased against the lower-performing students, the bias will be signed positive. If another item is found to be biased against the higher-performing students it will be signed negative. If these two biases on different items are of equal value they will, when aggregated, offset each other. This approach will allow the obtained result for aggregated item bias to reflect the degree and direction of total item bias in the test. Thus, the sum of all of the IBs will be designated AIB.

Aggregating the Item Components

For the purposes of this study, any bias found in an item for subgroups matched at a total test score below the cut-score was assumed to be bias against the subgroup with the lower performance on that item. In other words, all bias was assumed to be negative bias. However, as noted above, restrictions in the range of phi will reduce the size of the computed item bias which will, to some degree, reduce the value of the aggregate of item bias. What would result is a worst-case estimate (attenuated by the restriction in the range of phi) for the bias in the test.

Given these two assumptions -- (a) students with equal test scores are equally competent and (b) item bias is against the subgroup with the lower item performance -- a study can be designed which compares the presented item bias procedure to a traditional approach to bias detection.

Overview of Experimental Study

The stated purpose of this study was to evaluate the stability of the presented bias detection procedure. For an item bias procedure to be useful it must be stable across applications to data sets that are considered equal in bias content. Stability is a relative term. To test the stability of the presented bias procedure (hereafter referred to as the AIB -- aggregate of item bias) it must be compared to some standard. This standard was respresented by applying an adapted chi-square bias detection procedure (described below) to the same data set as the AIB and comparing the results. This adaptation of the chi-square represents the traditional approach to bias detection. One of the logical arguments for the usefulness of phi in the estimation of the AIB was its adjustment for the number of observations included in the calculations (N). Chi-square is noted for its sensitivity to differences in N. Therefore, the study compared the stability of the AIB and the adapted chi-square across different-sized data sets (i.e., different-sized subgroups of students), resulting in the critical comparison of stability when the size of the samples are systematically varied.

The determination of the sizes of these data sets was made within the perspective of a given data set. The data set used in this study was obtained from the Dallas Independent School District (DISD).

Following are descriptions of the instrument used to obtain students test performance data, the students used as subjects in this study, the process used to determine the sample sizes for the study, and the two bias detection procedures.

Instrumentation

The instrument used in this study was the 1980 administration of the Basic Objectives Assessment Test (BOAT), a minimum-competency test used as a graduation requirement in the Dallas Independent School District (DISD). The test contains 125 four-option multiple choice questions. Students in the school district must correctly answer at least 87 of the 125 test items on the BOAT to pass. Students who do not pass the test are not allowed to graduate from high school. In the event that a student does not pass the BOAT by the senior year in high school, they must take a remedial course which has as its final

examination, the BOAT. If a student fails to pass the BOAT, further opportunities are made available. Even after the student has left the school system, opportunities are made available to take the test and to be issued a diploma (given that all other requirements are met).

The BOAT was written by employees of the DISD. The test was written based upon a set of objectives that were developed by the DISD and evaluated by persons in the Dallas community. Once the objectives were determined, items were written and administered to students in the DISD. Student performance was reviewed and the items revised. The test was compiled and administered to one thousand randomly selected students. Performance from this sample was reviewed and the final test was approved by the DISD Board of Education as were the requirements indicated above.

Results from the first real administration of the BOAT were the data used in this study. The study was restricted to eighth-grade students because eighth-grade students of 1980 were the first class who were required to pass the test (at some time prior to twelfth grade) as a

graduation requirement. Students in higher grades were also given the test, but their graduation was not contingent on passing the BOAT. It was thought that only those students who were required to pass the test would be motivated to do well on the test. This lack of motivation in students in grades higher than eighth grade was thought to possibly impact student performance.

Items on the BOAT are divided into seven functional areas: consumerism (25 questions), community (30 questions), medical (15 questions), home (20 questions), employment (10 questions), government (15 questions), and information sources (10 questions).

Information about population performance and reliability indices computed on the BOAT and obtained from a DISD report (Arrasmith, 1979) are displayed in Table 1.

Table 1
Reliability, Validity, Central Tendency
and Dispersion of the BOAT

Reliability

Kappa Coefficient ----------	+ 0.90
Kuder Richardson (KR-20) ---	+ 0.90
Test-Retest ----------	+ 0.90

Validity

Correlation of BOAT to ITBS Reading Comprehension ---	+ 0.60
Correlation of BOAT to ITBS Mathematics ----------	+ 0.46

Central Tendency and Dispersion

Subpopulation	Means Score	Standard Deviation
White Students	93.75	21.67
Black Students	70.00	26.23
Black and White combined	80.54	24.28

The mean scores of the Black and White students indicates a clear difference between the measured competencies of these subpopulations. An average Black students scored one standard deviation below an average White students. This difference may indicate the possibililty of some bias in the test against Black students.

The indices of reliability on the test, indicate that the test is highly stable over administrations (test-retest) and internally consistent (KR-20) even though it covers a wide range of objectives. In fact the indicators are precisely the same numbers. Reliability is an important aspect of a test. If the measures of reliability were low, scores obtained on students taking the test would be suspect in that the true scores might vary greatly from the observed scores. The two reliability indicators (test-retest and KR-20) refer to the test score in the classical norm-referenced sense. The Kappa indicates the degree of consistency of decision based on the test score. This too is an indicator of test-retest reliability. If this index were low, the test would be unstable in its ability to make consistent judgement of competence. The high stability in decision making is relevant in this study since scores just below the cut-off will be used.

The correlations of the BOAT with the Iowa Tests of Basic Skills (ITBS) indicate that the BOAT correlates moderately with the ITBS reading and mathematics subtests. Such low correlations are not damaging to the validity of this application. The ITBS is a norm-referenced test of

academic skills; thus it should have relatively high variance. A minimum-competency test such as the BOAT is not constructed in a way which would maximize variance and should not correlate very highly. The higher correlation with the reading is likely a function of the reading intensive nature of the objectives which are measured by the BOAT.

Subjects

The subjects for this study were eighth-grade students in the DISD. Those included in this study were students whose computer files at the DISD contained total test score for the 1980 BOAT, item performance on the 1980 BOAT, and the race of the student indicating Black or White. A total of 19,849 eighth-grade students had complete computer files.

A review of the DISD records indicates the following numbers of eighth-grade students by race at total test scores of 85 and 86:

Total Test Score	N of Black Students	N of White Students	Total
86	158	91	249
85	152	95	247

Formation of Subgroups

To investigate the stability of the AIB and chi-square procedures across different sizes of student subgroups, several factors must be taken into account. Computations of the chi-square and phi are unstable when too small a sample size is used. Also the computer program used to compute the values for the chi-squares and phis in this study (Nie, Hull, Jenkins, Steinbrenner & Bent, 1970) requires a minimum of thirty-one (31) subjects. The size of the subgroups must vary to allow for comparison of sensitivity of the bias techniques to subgroup sizes and a sufficient number of subgroups must be formed to allow for comparison of the stability of the bias techniques. Also there is a limited number of students available from the population who have test scores within a given range.

Since the design of the procedure requires that the subgroups be of sufficient size and the number of students at each score point is limited, it was necessary to include within separate analyses students with scores other than one below the cut-score (86). Using students with scores more than two from the cut-score might have threatened the assumption that the study was run with students just short of acceptable competence and, hence, of equal ability. Therefore, only those students with total scores of 86 or 85 on the BOAT were included in the study. However, to preserve the contention that the students are of equal ability, the study analyzed the data in such a way that students at different score points were not included in the same analysis.

Since the racial subgroups must be balanced, the number of students at each total score is limited. There are 91 White students with a total score of 86; therefore, there can only be 91 Black students with total scores of 86 included in the study. For the same reason, only 95 Black students with a total score of 85 can be included. Thus, the study used a total sample of 372 students: 91 Black students and 91 White students with

test scores of 86, and 95 Black students and 95 White students with test scores of 85. At both score points the Black students were randomly selected for inclusion in the study. Given these restrictions and this pool of students, seven subgroup were constructed:

For students with a total score of 86:
Four subgroups of 32 students (16 Black and 16 White)
One subgroup of 54 students (27 Black and 27 White)

For students with a total score of 85:
One subgroup of 54 students (27 Black and 27 White)
One subgroup of 136 students (68 Black and 68 White)

This formation of subgroup pairs was chosen to allow for a reasonable number of comparisons across subgroups that varied in size to a reasonable degree. More subgroups could have been formed but the variation in subgroup size would have decreased. Fewer subgroups could have been formed to allow for greater variation of subgroup size but fewer estimates would have been generated for the bias techniques. The comparison of the stability of the bias techniques is enhanced with a greater number of estimates.

Description of the AIB

For each item on the BOAT, a phi was computed correlating the demographic characteristic with item performance. The computed phi was squared and then multiplied by the variance for that item. The result of that computation was the estimated bias in that item (IB) and considered to be biased against that demographic subgroup with the lower performance on that item. The assignment of bias to the proper demographic subgroup was accomplished by signing (positive/negative) the obtained IB for each item. If the White subgroup was the lower-performing subgroup on that item, the IB was signed negative. If the Black subgroup was the lower-performing subgroup on that item, the IB was signed positive. The estimate for the AIB was obtained by summing the IBs for all items on the BOAT. The obtained AIB estimate is thought to reflect both the direction and degree of bias contained in the items in the test. If the estimate of the AIB was positive, the test was considered biased against Black students. If the estimate of the AIB was negative, the test was considered biased against White students. If the estimate for the AIB was zero, the test was considered unbiased with respect to the race of the

student.

Description of the Adapted Chi-Square

The chi-square item bias procedure used in this study is similar to the procedure used by Scheuneman (1979), with modifications to avoid the criticisms of Camilli (1980) and Shephard et al. (1980). These modifications were: computation of the chi-square on subgroups of students matched by test score (rather than subgroups of students from a range of test scores), balancing of the number of students in each demographic subgroup (rather than use racial subgroups of unequal size), and the inclusion of incorrect responses in the chi-square (rather than using only correct responses). For each item on the BOAT a chi-square was computed using the same two-by-two frequency display used for the phi calculations in the AIB. The probability estimate (significance) of the calculated chi-square was determined. An item was considered biased if the probability estimate was less than or equal to .05 (Shephard et al., 1980). As with the AIB procedure, the assignment of bias was based on the performance of the subgroups on the item in question. If, on an item determined to be biased (chi-square probability

equal to or less than .05), the White subgroup was the lower-performing subgroup, the item was considered biased against that subgroup and the item was assigned a value of negative one (-1) to indicate that bias. If, on that item, the Black subgroup was the lower-performing subgroup, the item was assigned a value of positive one (+1) to indicate that the bias in the item is against that subgroup. For items not found to be biased (significance of more than .05 derived from the chi-square), the item was assigned a value of zero (0) indicating no bias. By adding together all of the bias values for the items (+1, 0 or -1) an estimate was obtained for the aggregate of item bias in the test based on the chi-square procedure. If the sum of these values was positive, the test was considered biased against Black students. If the sum of these values was negative, the test was considered biased against White students. If the sum of these values was zero, the test was considered unbiased (neutral) with respect to the race of the student.

This procedure is referred to as the SSTD (sum of signed trichotomous determinations). This name reflects the difference between it and the AIB procedure. The SSTD reacts discretely (there are no values other than whole

numbers) in a trichotomous fashion (+1, 0 or -1) and is subject to a determination of the presence of bias (the significance level of the chi-square is investigated to make this determination). This process involves dichotomous decisions (biased/not biased) but responds in a trichotomous fashion due to the direction of bias (positive, negative or zero) when the item has been determined to be biased. The results from the inspections of the individual items are referred to as TDs.

Comparison of Procedures

The results of the study were the estimates for the AIB and the SSTD for each of the seven subgroups (five at a total score of 86 and two at a total score of 85). The expected values of the AIBs are equal for each of the five subgroups of students with test scores of 86. If the assumption holds that students with test scores of 86 and 85 are of equivalent competence, then the expected values for the AIB on the two subgroups at a total score of 85 are equal to the AIBs for students at 86. The same is true for the SSTD -- the obtained values for the SSTDs are expected to be equal for all subgroups of students at each

of the score points and perhaps across score points.

To compare the stability of the AIB and the SSTD, the variances of the results of the two procedures on the seven subgroups were calculated. These variances are thought to reflect the stability of the two indices of bias -- the lower the variance, the more stable the index.

As previously stated, the range for the IBs is less than +1 to -1. This restriction in range will also restrict the variances of the AIB. In order to properly compare the stability of the two procedures, a modification of their ranges will be necessary. The range of IB and AIB are not equal to those of the TD and SSTD. This is due to the restriction of the range of phi by the distribution of scores within the two-by-two matrix of student performance. When the p-value for the item differs from .50, the range of phi will not be +1 to -1 (Cureton, 1959). Also, since the value of phi is squared and then multiplied by the variance of the item to obtain the estimate of IB, the size of the item variance will also restrict the value of IB. In the current study, the score chosen for analysis is 86 items correct out of 125 items on the test. This results in an average p-value of

.688. Given that the average or expected p-value for each of the items on the test is .688, the largest expected value for phi is .7214. This value is obtained by assigning one subgroup a p-value of .688 and the other subgroup a p-value of .000. This is the most extreme difference that is likely to occur on any item on the BOAT for these students. When the phi is computed for such an item, the result is .7214. The variance of an item with a p-value of .688 is .2215 (.688 x .322). If these values are placed into the computational formula for IB, the result is .1153. Since the range of the TDs is +1 to -1, one could expect that the variance for the SSTDs will be larger than the variance for the AIBs. The expectation for the size of this difference in variances at the item component level may be modified by considering the IB as the TD multipled by .1153. That is, the TD can be one while the IB can be expected to be only .1153. Consequently the variance of a variable composed of IBs is multiplied by .1153 to the second power. Hence the AIB is expected to have a much smaller variance. This difference can be modified by taking the inverse of .1153 to the second power or 75.352.

This means that in order to adequately investigate the difference in variance between these two measures, the value of the variance for the AIBs should be multiplied by 75.352. Although this has the appearance of the adjustments in phi that were described by Cureton (1959) in the calculation of phi-max, this adjustment is only one of size of variance and is done to accomodate the expected values of the variances of the two indices, and does not affect the interpretation of the results of the individual item analyses. It is important to note that for this analysis the covariances of the items are ignored, and the items are considered to be independent. The impact of the covariance of items on the aggregate of item bias is not known. The investigation of the impact of this covariance is beyond the scope of this study. In light of these factors the hypothesis statement for this study is:

HO: variance of SSTD = variance of AIB x 75.352

Since the SSTD represents a currently accepted procedure for the investigation of bias, the AIB must be proven more stable than the SSTD to recommend it as a bias investigation technique. To test these two bias techniques for stability, an F statistic will be

calculated for the ratio of the variances of the two bias detection techniques (Li, 1964):

$$F = \frac{\text{variance of SSTD}}{\text{variance of AIB} \times 75.352}$$

The variance of the SSTD is made the numerator of the ratio to reflect the null hypothesis; in the absence of evidence to support AIB over SSTD, the SSTD will be preferred. Given the ratio of the variances, the only way that a significant F can be obtained is for the variance of AIB to be significantly smaller than the variance of the SSTD. This would indicate the AIB as the more stable of the two procedures.

Summary

A procedure (AIB) was described which computes a correlation coefficient (phi) for students divided by race to item performance. The coefficient is squared and multiplied by the item variance to obtain a value which is thought to reflect the degree of bias in the item. These values are signed and then aggregated across all items on a minimum-competency test to obtain an estimate for the bias in the test. Another procedure (SSTD) was described which was an adaptation of the chi-square item bias detection procedure. A study was described that compared the stability of the AIB to that of the SSTD with the F statistic.

CHAPTER IV

RESULTS OF THE STUDY

The results of this study are described in three sections. The first section attempts to answer the primary question of this study -- is the AIB a more stable procedure than the SSTD when compared across data sets which vary by numbers of subjects? Subsequent sections investigate other aspects of the data and investigate certain assumptions made. These subsequent analyses hinge upon the results of the comparison of the stabilities of the two bias detection procedures.

Results of the Primary Analysis

Each of the subgroups upon which the SSTDs and AIBs were computed contained a balanced number of Black and White students, each with the same total score on the

test. The number of students in these subgroups was varied to test the sensitivity of the item bias procedures to differences in subgroup size, a condition known to disrupt the chi-square statistic used in the computation of the SSTD. The results of the calculations of the SSTDs and AIBs are displayed in Table 2.

TABLE 2
Results of SSTD and AIB Calculations

Subgroup	1	2	3	4	5	6	7
Test Score	86	86	85	85	85	85	85
n of Subgroup	54	134	32	32	32	32	54
n of Biased Items	6	16	12	9	5	5	4
SSTD	2	4	2	1	1	1	0
AIB*	.028	.011	.053	-.002	.000	.008	.004

*AIB rounded to three decimal places; accuracy to five decimal places retained in all calculations.

Since the subgroups were randomly assigned, any bias in the test would be expected to be consistent across subgroups. Therefore, the expectation was for there to be equal estimates for bias within each of these subgroups. Any difference between SSTDs computed across the seven subgroups would be seen as a sign of instability in the procedure. The same is true for noted differences between estimates for AIBs across the seven subgroups. This instability has been expressed as the variances of the two item bias procedures. The greater the variance, the greater the assumed instability of the procedure. Analyzing the ratio of the variances of the two procedures, with the F statistic, will provide an estimate of whether the two variances are equal. The test of the hypothesis of equality was made using an alpha of .05 for the comparison. Due to restrictions of the range of the phi computed for each IB in the AIB, the expected values of the AIB is less than the expected values for the SSTD. To adjust for this difference in range, the value of the variance of the AIB was multiplied by 75.35 . The result is displayed in Table 3.

Table 3
F-Test of the Ratio of SSTD to AIB

Variance of SSTD / Variance of AIB x 75.352 = F
1.6190476 / 0.0003853 x 75.352 = F
1.6190476 / 0.029033 = F
Degrees of Freedom --- 6 and 6
55.76577 = F
Probability = less than .0001

The value obtained for F (55.76577 with 6 and 6 degrees of freedom) is likely to occur only once in ten thousand calculations when there is in fact no difference between these two variances (i.e., when the null hypothesis is true). Such a low probability leads to the rejection of the notion that the variances of the indices are equal. This result supports the contention that the AIB is the more stable of these two item bias procedures. Note that the variance of the AIB, even when moderated, is substantially less than the variance of the SSTD.

Since these results support the possibility that the AIB is a useful procedure of bias detection, additional study is warranted. Within the limitations of this study, certain aspects of the results can be inspected to determine further the efficacy of the AIB procedure.

Secondary Analysis

The primary analysis in this study was the computation of SSTDs and AIBs on the seven student subgroups and the subsequent computation of the ratio of the variances of the two item bias procedures. These seven subgroups were at two total score points: students in subgroups one and two had total test scores of 86, students in subgroups three through seven had total test scores of 85. The results computed on the test performance of these seven subgroups were combined in the overall comparison of the two item bias procedures. The combination of the results at these two score points was based on the assumption that students at the two score points were equivalent in their performance on the test and should therefore be equivalently sensitive to bias in the test. Also the sizes of the groups were varied to test whether the AIB would be as sensitive to the number

of subjects used in the calculation as the chi-square procedure.

As a check to see if the results at these two total score points were equivalent, a secondary study was run. A test of equivalence would be a mean comparison of AIB estimated at a score of 85 to AIB estimated at a score of 86, and a comparison of SSTD at 86 to SSTD at 85. This comparison is within procedure, not across procedures, as in the primary analysis. In this secondary study, the same sample of students was randomly assigned to fourteen subgroups, seven at each total score point. Each of these fourteen subgroups contained twelve Black and twelve White students.

Thus the factor of different groups sizes would not be present in the calculations of AIB and SSTD on these groups. Within each of these subgroups all students had identical total scores on the BOAT: subgroups one through seven contained only those students with scores of 86, subgroups eight through fourteen contained only those students with scores of 85. SSTDs and AIBs were computed for each of these fourteen subgroups, the results of these analyses are shown in Table 4.

Table 4
Result of SSTD and AIB
for Subgroups with Balanced N's

	Subgroup #													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Test Score	86	86	86	86	86	86	86	87	87	87	87	87	87	87
n of Subgroup	24	24	24	24	24	24	24	24	24	24	24	24	24	24
n of Biased Items	5	5	5	5	7	10	7	6	5	11	7	6	7	3
SSTD	-1	3	5	-1	-1	2	1	0	1	-3	3	0	-3	-1
AIB*	.007	.128	.122	.000	-.031	.006	.076	.052	.042	-.150	.066	.007	-.066	-.021

*AIB rounded to three decimal places; accuracy to five decimal places retained in all calculations.

To determine if the results at the two score points (85 and 86) are equivalent in measured bias, a t-test was run on the SSTDs and AIBs computed for the groups at the two score points. If the results of the t-test indicated that there was a significant difference between the SSTDs at 85 and the SSTDs at 86, or between the AIBs at 85 and the AIBs at 86, their combination in the primary analysis might not be suggested. The results of the t-tests are displayed in Table 5.

Table 5
t-Test of SSTDs and AIBs at the two Score Points

Group	Size	Mean	Variance	t Value	Degrees of Freedom	Probab.
SSTD 85	7	1.14	5.476			
SSTD 86	7	-0.42	4.619	1.309	13	0.215
AIB 85	7	0.05	0.004			
AIB 86	7	-0.01	0.006	1.676	13	0.120

Using an alpha of .05 the null hypothesis of no difference cannot be rejected. Thus, these t-tests would support the notion that the results at the two total scores are equivalent or at least not significantly different, and

that their combination in the primary analysis was reasonable, in that neither procedure produced different results across the two score points.

Testable Assumptions

The data obtained in the study can be inspected under certain assumptions. By inspecting the outcomes of analyses performed under these assumptions, information can be obtained which may be useful in the evaluation of the AIB and the methodology used in this study. The findings of these inspections can be compared to the expected outcomes and used to investigate the efficacy of the AIB and SSTD procedures.

The first assumption deals with the signs (+, -) of the bias estimates. The sign of the bias estimates indicates the direction of the bias -- a positive bias estimate indicates that the test was biased against Black students; a negative bias estimate indicates that the test was biased against White students; a zero bias estimate indicates that the test was neutral with respect to these two subpopulations. Since the subjects were randomly assigned to the subgroups, any bias in the test

should be consistent across subgroups. Although some variation due to sample fluctuations can be expected across subgroups, the direction of the bias should remain consistent. In reviewing the results of the computations of SSTD and AIB (see Table 2) it can be seen that only one of the fourteen bias estimates (AIB for group number 4 in Table 2) indicated that the direction of the bias was negative, two indicated that the test was neutral (AIB for group number 5 and SSTD for group number 7 in Table 2). The remaining eleven bias estimates indicated the direction of the bias as positive. This overall consistency in direction of bias supports both the SSTD and the AIB as reliable bias estimates. The results fom the secondary study (see Table 4) show four negative estimates for the AIB and six negative estimates for the SSTD. This decrease in agreement with the population data may be due to the small sample sizes used in the secondary study.

The second assumption concerns the number of biased items found in the test. This number of bias items is derived from the chi-square analyses from the SSTD calculations. If an item was found to exceed the .05 level on the chi-square analysis, it was considered to be

biased. The number of biased items found for each of the subgroups is displayed in Table 2. Since the chi-square is more sensitive to item bias when more subjects are included in the calculations, the expectation would be for the number of biased items detected to increase with the number of subjects in the subgroups. (The chi-squares on the larger groups are able to detect amounts of bias that are undetectable in analyses with smaller subgroups.) The results are inconclusive; the greatest number of biased items was found in the subgroup containing the greatest number of students (sixteen biased items in subgroup two in Table 2). However, the average number of biased items for the subgroups with 54 students (5.0 items) was lower than the average number of biased items for subgroups containing 34 students (7.75 items). These results indicate the possibility that smaller subgroup sizes may be somewhat unstable in estimated bias due to random fluctuation in the sampled students. This is supported by the secondary study (see table 4) where the sample size was consistently 24, the range of biased items was from 3 to 11, and the mode was 5.

The third assumption deals with population data. If a test is biased against one racial subgroup, lower mean scores can be expected for that racial subgroup when compared to other racial subgroups. It is unlikely that a test which shows racial subgroup A to be superior to racial subgroup B is in fact biased against racial subgroup A. In the perspective of this study, if Black students were indicated to be superior, overall, to White students on the BOAT, it would be difficult to explain were the test shown to be biased against Blacks.

Total test scores for the two racial subgroups were inspected. For the entire population of students taking the BOAT (approximately 20,000 students) the median scores for these two subpopulations are displayed in Table 6.

TABLE 6
Median Scores of
White and Black Students on the BOAT

Median score of all White students ------	100
Median score of all Black students -----	75

From these results and the mean for each group given in Table 1 it would seem likely that the BOAT was either biased against Black students or unbiased and reflecting true differences in competence between the two subpopulations. Estimates indicating that the BOAT was biased against White students would be suspect.

The mean of the SSTDs for the seven primary groups was computed, as was the mean for the AIBs. A positive mean would indicate that the measure (either SSTD or AIB) showed the test to be biased against Black students. A mean of zero would indicate that the test was unbiased. A negative mean would indicate that the test was biased against White students. It is only this last result (negative mean) that would contradict the information obtained from the population scores. The mean of the two item bias measures (based on the primary anlysis only) were:

Mean of SSTD ---- +1.51714

Mean of AIB ----- +0.0287

These results are consistent with the differences observed between the total test scores of the two racial subgroups,

indicating that the BOAT may be biased to some degree against Black students. These results add to the validity of both of these procedures.

The last assumption to be investigated deals with the correlation of the two item bias measures. If the AIB is in fact more stable than the SSTD across data sets of different sizes, the AIB should correlate better with the SSTD across data sets that do not vary in size. When data sets are the same size, fluctuations in the SSTD due to the differences in size should not be present. Within the perspective of this study, the SSTD would be expected to correlate higher with the AIB in the secondary analysis than in the primary analysis. In the secondary analysis, the subgroups were all of the same size (see Table 4). In the primary analysis, the sizes of the subgroups varied from 32 to 154 (see Table 2). For each of the analyses (primary and secondary) the correlation between the SSTD and AIB was calculated. The following results were obtained:

Correlation for the primary analysis --- 0.3451 (5 df)

Correlation for the secondary analysis - 0.9072 (12 df)

Glass and Stanley (1970) list the critical values for correlation coefficients as .811 for coefficients with five degrees of freedom and .707 for twelve degrees of freedom (alpha of .05 for both values). The value obtained for the secondary analysis indicates significance. The value obtained for the primary study does not indicate significant correlation. These results support the hypothesis and lend further credence to the use of the AIB.

Summary

When the AIB and SSTD procedures are applied across a number of subgroups that vary in the number of subjects, the AIB appears to the more stable of the two measures. Both the AIB and SSTD procedures yield results which are consistent with population data. The comparison of results across the two score points indicates that it is reasonable to combine these data in the comparisons of the two item bias procedures. When the two procedures are applied to data sets that do not vary in the number of subjects, the differences between the two procedures diminish.

CHAPTER V

DISCUSSION

This chapter is comprised of three sections: the first section discusses the results of the study; the second section discusses the limitations of the study; and the third section suggests area for further research.

Discussion of Results

Within the perspective of the study, the results are quite supportive. With the exception of the number of biased items as a function of subgroup size, the SSTD and AIB performed as expected. If the experimental procedure is conceptually sound, the size of the obtained F leads to a clear decision to reject the null hypothesis: The AIB is the more stable of the two procedures. The results of the secondary study support the assumption of equivalence of the data at the two total scores (85 and 86).

The stability of both procedures (SSTD and AIB) over the subgroups furthers the support for the approach to bias in MCTs used in this study. The results, when compared to population data, suggest that the item bias procedures may give information useful in assessing test bias. Having a procedure for estimating test bias which requires no external criterion offers the possibility of considerably decreasing the effort and expense in validating the lack of bias in a minimum-competency test. In addition to the expense and effort, test bias procedures involving comparisons of test scores to some external criterion are limited by the reliability of the external criterion and the validity of its application. This can be seen in the low validity coefficients between the BOAT and the ITBS. Had test bias in the BOAT been assessed in reference to ITBS scores, more than sixty percent of the test score variance in the BOAT would remain unaccounted for in the ITBS scores (the coefficient of determination for the ITBS reading scores is .36 and for the mathematics score it is .2116).

Limitations of the Study

The description of the limits of the study is not a critique of the study. To describe the limits is to define the boundaries of the study. One study in an area not otherwise investigated can offer little other than direction for further research. Prior to further research, little can be said of the usefulness of the AIB procedure outside of the setting to which it was applied in this study.

The study used the 1980 administration of the Basic Objectives Assessment Test (BOAT), a minimum-competency test developed and used as a graduation requirement by the Dallas Independent School District. The subjects were limited to those Black and White eigth-grade students with test scores of 85 and 86 on the BOAT. The impact of changing any of these factors is not known. Prior to further investigation the AIB should be viewed as an experimental procedure.

Applications of the AIB should consider certain other limitations. The demanding assumption of equal competence detailed in the Methodology Chapter can only apply to minimum-competency tests. The AIB can therefore only be reasonably applied to minimum-competency tests. Test bias

is not limited to the aggregate of item biases. Factors other than independent item bias can cause a test to be biased. The impact of the covariance of the item biases on the aggregate of item biases is not known. Additionally, evironmental factors such as the quality of instructions for test taking, the quality and range of sample questions, time constraints, and other factors may differentially impact the performance of one subpopulation. Also, the phi is limited in the estimation of bias in test items with p-values that differ from .5. This suggests that the AIB may be most useful on minimum-competency tests with cut-scores close to fifty percent of the total number of items on the test.

Interpretations of the results of the AIB is limited by the lack of a complete understanding of what the procedure measures and the impact of item covariances. Prior to further study no attempt can be made as to the interpretation of the results of the AIB procedure.

Suggestions for Further Research

For the AIB to be of any use to the field of education, further investigation will be required. The best validations are obtained through systematic replications and applications to other situations. Several extensions of the current study can be suggested.

One of the most important areas for further study is the adjustment factor used to modify the two variances prior to the computation of the F of the ratio of the variances. The use of the factor of 75.35 was based on an analysis of the components in the two procedures. A more direct method of determining the two variances can be described. Students from the sampled population could be randomly assigned to subgroups. These subgroups would not be expected to differ in composition (e.g., race). Since no difference is expected between such randomly assigned groups, any variance in bias estimates computed can be considered random variation in the measure and viewed as "noise" in the indices. The ratio of the variances for the SSTD and AIB computed for these subgroups directly reflect the differences in the ranges for these two measures. If this obtained ratio differs from 75.35, then that adjustment

factor could be used to recompute the F of the variances.

Another study that would be of use in the investigation of the AIB would be the application of the AIB procedure to a data set known to contain bias. This data set could be one that has been found to be biased in previous investigation, or one in which the bias was induced in the data set by simulation. Particularly useful results could be gathered if the degree of bias was manipulated, since the AIB considers degrees of bias.

Investigation of the impact of item covariance on the AIB would also be useful. This investigation would most likely be best done with data sets where item bias and item covariances are simulated. Such an investigation might lead to an understanding of what the AIB measures and how it may be properly interpreted.

The AIB is not limited to the investigation of bias at test scores below the cut-score. By reversing the sign (+, - or 0) assignment of bias to the subgroups, the AIB can be used to investigate positive bias for those students above the cut-score. This would be useful in the investigation of the impact of test bias in the passing of students who might

not have passed the test had the bias not been contained in the items. By combining the results of AIB calculations above and below the cut-score, a band of unsurety, much like the standard error band in norm-referenced testing, can be determined where interpretation of test scores might be impacted by the aggregated item biases. This would be useful in validating criteria for passing the test.

In any event, the results of this study are clearly supportive of the AIB. It appears that it may be possible to estimate some component of test bias in minimum-competency tests through the assessment of component item biases. The usefulness of such a procedure in the validation of minimum-competency tests may make the AIB an attractive possibility.

BIBLIOGRAPHY

Angoff, W. H., & Ford, S. F. Item-race interaction on a test of scholastic aptitude. Journal of Educational Measurement, 1973, 10, 95-106.

Arrasmith, D. Initial implementation and technical description of the basic objectives assessment tests. Dallas, Texas: Dallas Independent School District, Department of Research, Evaluation and Information Systems, May 1979.

Berk,R. A., Criterion-referenced testing: state of the art. Baltimore: The Johns Hopkins University Press, 1980.

Bunda, M. A. & Sanders, J. R. (Eds.) Practices and problems in competency-based measurement. Washington, D. C.: National Council on Measurement in Education, 1979.

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

Carrol, John, B. The nature of the data, or how to choose a correlation coefficient. Psychometrika, 1961, 26, 347-72.

Cole, N. Bias in selection. Journal of Educational Measurement, 1973, 10, 237-255.

Cronbach, L. J. Test validation. In R. L. Thorndike (Ed.), Educational Measurement (2nd Ed.). Washington, D.C.: American Council on Education, 1971.

Cureton, E. Note on phi/phi max. Psychometrika, 1959, 24, 89-91.

Ebel, R. L. Evaluation and educational objectives. Journal of Educational Measurement, 1973, 10, 273-279.

Glaser, R. Instructional technology and the measurement of learning outcomes. American Psychologist, 1963, 18, 519-521.

Glass, G. V. & Stanley, J. C. Statistical methods in educational psychology. Englewood Cliffs, New Jersey: Prentice-Hall, 1970.

Gray, W. M. A comparison of Piagetian theory and criterion- referenced measurement. Review of Educational Research, 1978, 48, 223-249.

Green, D. R. & Draper, J. F. Exploratory studies of bias in achievement tests. Paper presented as the Annual Meeting of the American Psychological Association, Honolulu, 1972.

Hambleton, R. K., Mills, C. N., Simon, R. A. & Livingston, S. Constructing and using criterion-referenced tests. Materials from the 1980 American Educational Research Association training session. Available through Northwest Regional Labs, Portland, Oregon.

Hambleton, R. K. & Novick, M. R. Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 1973, 10, 159-170.

Hambleton, R., Swaminathan, H., Algina, J. & Coulson, D. Criterion-referenced testing and measurement. Review of Educational Research, 1978, 48, 1-47.

Hopkins, K. & Glass, G. Basic statistics for the behavioral sciences. Engelwood Cliffs, New Jersey: Prentice-Hall, 1978.

Ironson, G. H. & Subkoviak, M. J. A comparison of several methods of assessing bias. Journal of Educational Measurement, 1979, 10, 209-225.

Jensen, A. Bias in mental testing. New York: Free Press, 1980.

Li, J. Statistical Inference I. Ann Arbor, Michigan: Edwards Brothers, Inc., 1964.

Linn, R. Fair test use in selection. Review of Educational Research, 1973, 43, 139-161.

Linn, R. Reliability. Chapter Five in Bunda & Sanders (Eds.); Practices and problems in competency-based measurement. Washington, D. C.: National Council on Measurement in Education, 1979.

Linn, R. & Harnish, D. Interaction between item content and group membership on achievement test scores. Paper presented to the Annual Meeting of the American Educational Research Association, San Francisco, California, 1979.

Lord, F. M. Practical applications of item characteristic curve theory. Journal of Educational Measurement, 1977, 14, 117-138.

Millman, J. Criterion-referenced measurement. In W. J. Popham (Ed.), Evaluation in education. Berkeley, California: McCutchan, 1974.

Nie, M., Hull, C., Jenkins, J., Steinbrenner, K. Bent, H. Statistical package for the social sciences (second edition). New York: McGraw Hill, 1970.

Popham, W. J. Well-crafted criterion-referenced tests. Educational Leadership, 1978, November, 91-95.

Popham, W. J. & Husek, T. R. Implications of criterion-referenced measurement. Journal of Educational Measurement, 1969, 6, 1-9.

Rudner, L. M. & Convey, J. J. An evaluation of select approaches for biased item identification. Paper presented at the annual meeting of the American Educational Research Association, Toronto, Canada, 1978.

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

Sanders, J. R. & Murray, S. L. Alternatives for achievement test. Educational Technology, 1976, 16, 17-23.

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

Shephard, L., Camilli, G. & Averill, M. Comparison of six procedures for detecting test item bias using both internal and external ability criteria. Paper presented to the Annual Meeting of the National Council on Measurement in Education, Boston, 1980.

Swaminathan, H., Hambleton, R. & Algina, J. Reliability of criterion-referenced tests: A decision-theoretic formulation. Journal of Educational Measurement, 1974, 11, 262-267.

Thorndike, R. Concepts of culture-fairness. Journal of Educational Measurement, 1971, 8, 63-70.

Wingersky, M. Test whether the item parameters estimated by LOGIST for two separate groups differ significantly. Princeton, New Jersey: Educational Testing Services, 1977.

Wright, B. D. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977, 14, 97-116.

Wright, B. D., Mead, R. J. & Draba, R. Detecting and correcting test item bias with a logistic response model. Research Memorandum 22, Statistical Laboratory, Department of Education, University of Chicago, 1976.