INFORMATION TO USERS

This reproduction was made from a copy of a document sent to us for microfilming. While the most advanced technology has been used to photograph and reproduce this document, the quality of the reproduction is heavily dependent upon the quality of the material submitted.

The following explanation of techniques is provided to help clarify markings or notations which may appear on this reproduction.

1. The sign or "target" for pages apparently lacking from the document photographed is "Missing Page(s)". If it was possible to obtain the missing page(s) or section, they are spliced into the film along with adjacent pages. This may have necessitated cutting through an image and duplicating adjacent pages to assure complete continuity.

2. When an image on the film is obliterated with a round black mark, it is an indication of either blurred copy because of movement during exposure, duplicate copy, or copyrighted materials that should not have been filmed. For blurred pages, a good image of the page can be found in the adjacent frame. If copyrighted materials were deleted, a target note will appear listing the pages in the adjacent frame.

3. When a map, drawing or chart, etc., is part of the material being photographed, a definite method of "sectioning" the material has been followed. It is customary to begin filming at the upper left hand corner of a large sheet and to continue from left to right in equal sections with small overlaps. If necessary, sectioning is continued again—beginning below the first row and continuing on until complete.

4. For illustrations that cannot be satisfactorily reproduced by xerographic means, photographic prints can be purchased at additional cost and inserted into your xerographic copy. These prints are available upon request from the Dissertations Customer Services Department.

5. Some pages in any document may have indistinct print. In all cases the best available copy has been filmed.

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

Schoener, John Edwin

A COMPARISON OF STATISTICAL AND JUDGMENTAL METHODS FOR IDENTIFYING ITEM BIAS

City University of New York PH.D. 1984

PLEASE NOTE:

In all cases this material has been filmed in the best possible way from the available copy. Problems encountered with this document have been identified here with a check mark __√__.

1. Glossy photographs or pages _____
2. Colored illustrations, paper or print _____
3. Photographs with dark background _____
4. Illustrations are poor copy _____
5. Pages with black marks, not original copy _____
6. Print shows through as there is text on both sides of page _____
7. Indistinct, broken or small print on several pages __✓__
8. Print exceeds margin requirements _____
9. Tightly bound copy with print lost in spine _____
10. Computer printout pages with indistinct print _____
11. Page(s) _____ lacking when material received, and not available from school or author.
12. Page(s) _____ seem to be missing in numbering only as text follows.
13. Two pages numbered _____. Text follows.
14. Curling and wrinkled pages _____
15. Other _____

A COMPARISON OF STATISTICAL AND JUDGMENTAL METHODS FOR IDENTIFYING ITEM BIAS

by

JOHN E. SCHOENER

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

1984

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

7/9/84
date

Chairman of Examining Committee

7/11/84
date

Executive Officer

Professor Alan L. Gross

Professor David Rindskopf

Professor Max Weiner
Supervisory Committee

The City University of New York

Abstract

A COMPARISON OF STATISTICAL AND JUDGMENTAL METHODS FOR IDENTIFYING ITEM BIAS

by

John E. Schoener

Adviser: Professor Alan L. Gross

The purpose of this study was to compare test items identified as biased using statistical and judgmental procedures. Several critical questions were investigated: Will test items identified as biased by reviewers from different subgroups of the population be related across subgroups? Will test items identified as biased by a statistical procedure for ethnic subgroups of students be related to those identified as biased for gender subgroups of students? Will statistical and judgmental methods for identifying biased items on a test agree? The affect of rescoring the test eliminating items identified as biased was examined and the correlations of the test and the rescored test with an external criterion were determined.

A criterion-referenced mathematics test was administered to 1064 high school students of both sexes and diverse ethnic backgrounds. The three-parameter latent

trait model was the statistical procedure used to detect biased items. The judgmental procedure consisted of a review of the test items by twenty-four judges who were knowledgeable about high school mathematics curricula. Judges used a structured rating form. Eight of the reviewers were black, eight of the reviewers were white, and eight of the reviewers were Hispanic. Within each group, half of the reviewers were male and half were female.

Agreement between statistical ratings among subgroups, agreement between judgmental ratings among subgroups and agreement between the statistical and judgmental procedures was assessed using the Kappa statistic. There was no significant agreement between statistical bias ratings for ethnic and gender subgroups. There was significant agreement on some of the indicators calculated to combine judges ratings, but not on others. There was no significant agreement between item bias detection methods. Rescoring the test, eliminating the items identified as biased by the statistical procedure, did not change the rank order of subgroups and the total group. Both the test and the rescored test had significant correlations with a standardized, norm-referenced mathematics test.

Table of Contents

List of Tables

List of Tables - Continued

List of Figures

Introduction

The issue of test fairness or test bias is not new (Cronbach, 1975), but the definition of bias has evolved rapidly over the past decade. Recent court decisions and federal legislation have made tests the object of close public scrutiny. Federal legislation (Federal Register, 1977) has called for "non-biased assessment" of all groups. The unequal representation of students from diverse cultural groups in programs for exceptional children has been attributed to bias in tests. Litigation, often involving ethnic or racial subgroup members and boards of education, has underscored this concern about fair tests and the fair use of tests (Larry P. et al. v. Riles, Diana et al. v. California State Board of Education).

Within this arena, it has been difficult for test makers and test users to counter these charges that pertain to sex, race, or ethnic bias. While a charge of bias may be relatively easily leveled against almost any assessment instrument, it is a far more thorny problem to obtain definite evidence to substantiate or refute these charges (Berk, 1980). Various statistical and judgmental methods have been proposed to detect item bias. However, there have been few studies to date in which these approaches have been

compared. The purpose of the proposed study is to compare statistical and judgmental methods for identifying biased items.

The terms "test bias," "cultural bias," and "cultural fairness" have been defined and used in many ways in the past ten years. Jensen (1980) has pointed out, however, that concerns about "cultural bias" began as early as 1905 when Binet and Simon acknowledged the problem in their new intelligence test. The renewed interest in the area is due, to a large extent, to growing public concern about testing in recent years.

School systems, the judicial system and private enterprise (test developers/publishers) have become complexly interlocked in the test fairness issue. Each of these institutions has a somewhat different stake in the use or misuse of tests, but each has been forced to confront the issue. In the past fifteen years, there have been numerous landmark court cases in which the plaintiff (often an individual or his/her parents) has contended that he or she has been discriminated against through the use of a test. The defendant (often a school system or an employer) has claimed that the test is being used to provide the best educational opportunity for the individual or to identify the best individual for a particular position (Hobson v. Hansen; Griggs et al. v. Duke Power Company). If the court deems a test unfair, and this has happened with increasing

frequency, the test can no longer be used in the situation in which it was ruled as biased. Thus, test publishers have a major interest in developing tests that will be considered fair in the public eye. Judges in these legal cases have used various criteria in making rulings about the fairness of tests: selection and placement decisions (based on statistical evidence), content or facial evidence of fairness, and the testimony of expert witnesses about material that may be considered to be offensive and/or unfamiliar to particular gender, ethnic or racial subgroups.

Psychometricians have attempted to address this problem by developing suitable selection and prediction models for the placement of individuals in school and employment settings. Predictive validity has been the central issue. A number of different statistical models have been described in the literature for assessing bias in the use of tests (Cleary, 1968; Thorndike, 1971; Darlington, 1971; and Gross and Su, 1975). These models, which stress fairness in selection, were reviewed by Peterson and Novick (1976). They all involve different selection philosophies and whether or not a test is biased for a particular use. These approaches may correct for what is considered unfairness according to one of these philosophies, but the use of these models in no way alters the characteristics of a test or items on a test.

In the last several years a new issue has emerged in the area of test bias: potential bias inherent within the items of a test. Investigators have begun to explore statistical models for identifying biased items. The concern here is with the content validity and construct validity of an instrument. Merz, Grosen, Groome and Groome (1979) have used the terms test bias and item bias to distinguish between bias on a test as a whole and bias inherent within the items of a test. Merz et. al. maintain that test bias is determined by whether or not similar predictions about the success of individuals in a particular setting are obtained when different groups are tested using the same instrument (see also Potthoff, 1966; Merz, 1978; Rudner, 1977; Merz and Rudner, 1978). Item bias, Merz et. al. contend, is a property of the items themselves and is not related to external criteria. Item bias should be examined during the test construction process and should be determined by an examination of test items in relation to total test score, item difficulties, or other criteria specific to the individual test.

In order to arrive at a conceptual definition of bias, the distinction between bias in the items of a test and bias in test use must be explored. Indeed, they are not separate concerns, but essential and related elements of the larger issue of fair testing. Shepard (1980) has stated, "Most authors define bias as a type of invalidity. Bias then is

taken to be an inherent feature of a test, while its opposite, validity, has always been considered to be a property of test use, not of the test itself" (p.2). Shepard has suggested, however, that:

> There is a validity continuum, anchored at one end by unbiased tests which measure what they were designed to measure and do equally well for all groups. Further along the continuum are tests that provide equal predictive validity in particular contexts. At the other end of the continuum are tests for which validity involves issues of social justice and values, as well as scientific arguments over what statistical model of fairness to apply and what the criterion will be (p.3).

Given the larger social context, statistical methods of identifying biased items or for adjusting for unfairness in test use, while important, are by themselves inadequate for addressing the broader theoretical issues of fair testing. Cronbach (1980) has also pointed out the need for a new look at validity in this area. With regard to validity, the question asked historically has been: "What does the test measure?" Cronbach has suggested that this question alone is no longer sufficient, but that we must also ask "Why should that be measured?"

Statistical models of selection are essential in providing answers to predictive validity questions. However, construct validity and content validity, as well as issues of social justice, must also be considered. In order to address these issues, responses to the items on a test across subgroups must be studied. Shepard (1980) has

contended that "unbiasedness in test items is best conceptualized as equal construct validity across subgroups and must be confirmed by verifying expected patterns of relationships." She has maintained that it is necessary to employ statistical analyses and reviews by expert judges and subgroup members to fully understand the concept or trait being investigated.

Statistical methods for identifying biased items provide an indication of whether the meaning of an individual item on a test is the same for all subgroups. However, statistical methods do not provide insight about possible reasons for different patterns of item responses, which may indicate bias for different subgroups. Reviewers who are knowledgeable about the test content and the examinee population and reviewers who are members of different racial, ethnic or gender subgroups can contribute to an understanding of different response patterns. In addition, Angoff (1980) has stated that "proper methods of test development require extensive and careful judgmental review of all items from the point of view of ethnic and sex bias, among others, before they are accepted for inclusion in a test." Statistical methods are a tool to be used by test developers and reviewers. Angoff (1980) has pointed out that "statistical methods are only supplemental to human judgment, certainly no substitute for it."

Thus, both statistical and judgmental item bias detection methods should be employed in establishing the construct validity and content validity of a test for different subgroups. However, few studies appear in the literature in which these procedures have been compared or combined in the test development or revision process. The purpose of the proposed investigation is to systematically utilize and compare statistical and judgmental item bias detection methods.

This dissertation is divided into eight sections: 1) a review of studies that examine different statistical methods for identifying biased items; 2) a review of the use of judgmental methods for the identification of item bias; 3) a discussion of new approaches to identifying biased items within tests; 4) a statement of the problem; 5) questions to be investigated; 6) method; 7) results; and 8) discussion.

Review of the Literature

Statistical Item Bias Methods

The four statistical models most frequently used for identifying bias inherent within items are: item discrimination indices, transformed item difficulties, chi-square (with three intervals, five intervals or multiple intervals), and item characteristics curves (ICC) (Green and Draper, 1972; Angoff and Ford, 1974; Scheuneman, 1975; Hambleton, 1978; Lord, 1980). Each of these models is reviewed briefly in this section. In addition, studies comparing the efficacy of these models will be discussed.

Bias methods based on item discrimination indices have involved the traditional correlation of item score and total test score. Green and Draper (1972) were among the first researchers to use point biserial correlations. Using this method differences between pairs of correlations for each item are computed for two groups at a time. If the comparison for a particular item exceeds a prespecified value, the item is considered to be biased.

The delta-plot method has been the most widely used bias identification method. It is based on transformed difficulties and has been cited by many investigations as one of the most viable methods for identifying item bias

(Raju, 1980; Burrill, 1980). Angoff (1980) has reviewed this method as it was employed by Angoff and Ford (1974). Using this method, item difficulties (p-values) are calculated for two different groups. The p-values are transformed to standardized values by calculating the normal deviates (z-scores) for each group independently where z is given by:

$$z = \frac{p_i - \bar{p}}{S_p}$$

where p_i = the percent of test-takers responding correctly to the i^{th} item;

$\bar{p}$ = the average percent correct for the entire test;

S_p = the standard deviation of the percent correct for the entire test.

In order to compare item difficulties for the two groups, the z scores are converted to delta values (Δ) using the formal $\Delta = 4z + 13$, where the mean of the Δ's is 13 and the standard deviation is 4. The arbitrarily chosen mean and standard deviation for this linear transformation are used in order to remove negative values. The pairs of deltas, one pair for each item on the test for each group, are then plotted on a bivariate graph. The delta-values for one group are read on the ordinate. If the two groups are drawn from similar populations, the scatter plot of these points will fall on a long narrow ellipse. Those points

that fall at some distance from the major axis of the ellipse are more difficult for one group than the other relative to the other items on the test. This method may, of course, be used with more than two groups and a multi-dimensional ellipsoid will be formed. Angoff (1975) has suggested that in the case of a two-group comparison, a line can be determined that will run through the major axis of the multidimensional ellipsoid and the perpendicular distance can be found between an item point in the hyper-space and the major axis line. With regard to the outlined points, Angoff (1980) has suggested, "A clinical review of the items that fall at some distance from the plot may reveal that the group for which they are inordinately difficult may not have had the same amount of exposure to the concepts measured by these items as did the other group. It is in this sense that the items are referred as "biased" items" (p.4).

Chi-square methods for the identification of biased items compare entire distributions of responses for the groups in question. In this way they differ from the single item-parameter methods which use only a difficulty or dis-crimination index. Scheuneman (1975) has presented the most frequently used chi-square method for identifying biased items. In providing an operational definition of her pro-cedures, she has stated: "An item is unbiased if, for all individuals having the same score on a homogeneous subtest

containing the item, the proportion of individuals getting the item correct is the same for each population group being considered" (1975; p.2). Scheuneman has divided each group into three to five intervals based on observed total score. She then compares the proportion of individuals within each (ability) interval responding correctly. If the proportion of correct responses to an item is the same within score intervals for both groups being considered, the item is considered unbiased. The probability that an item is unbiased is estimated using a modified chi-square technique. For each cell, the expected values (E_{ij}) are obtained by multiplying the proportion of correct responses to an item for all those in either group within a level of ability by the number of respondents within a single group and ability level. That is:

$$E_{ij} = \frac{O \cdot j}{N \cdot j} \quad N_{ij}$$

Where: $O \cdot j$ = the number of examinees at ability level j who respond correctly to the item.

$N \cdot j$ = the total number of examinees at ability level j who respond to the item.

N_{ij} = the total number of examinees in group i and at ability level j.

Ability Level (j)

Group I	E_{ij}		
Group II			

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(E_{ij} - O_{ij})^2}{E_{ij}}$$

The chi-square is calculated by adding the sum of the squared differences between expected and observed frequencies divided by the expected frequencies for groups.

The use of latent trait theory or item characteristic curves in the identification of biased items has become prevalent in the past several years (Wright, Mead, and Draba, 1978; Hambleton et al., 1978; Lord, 1980). Using latent trait theory, observable scores to items on a test are related to an unobservable trait or ability (Θ) which determines these scores. Item characteristic curves (ICCs) describe the relationship between an examinee's ability (Θ) and the probability of a correct response to a test item (g) (Merz et al., 1978). This is a conditional distribution where the probability of a correct response to an item is expressed as $P(u_g = 1/ \Theta_i)$. An example of a hypothetical ICC is presented in Figure 1. A major development in this area has been the application of maximum likelihood methods to calculate the estimates of item and ability parameters when the ICCs are assumed to have a certain parametric form (Lord, 1980). In discussing the use of ICCs, Hambleton et al. (1978) have described two assumptions that must be met: unidimensionality and local independence. Unidimensionality refers to the assumption that a test measures a single

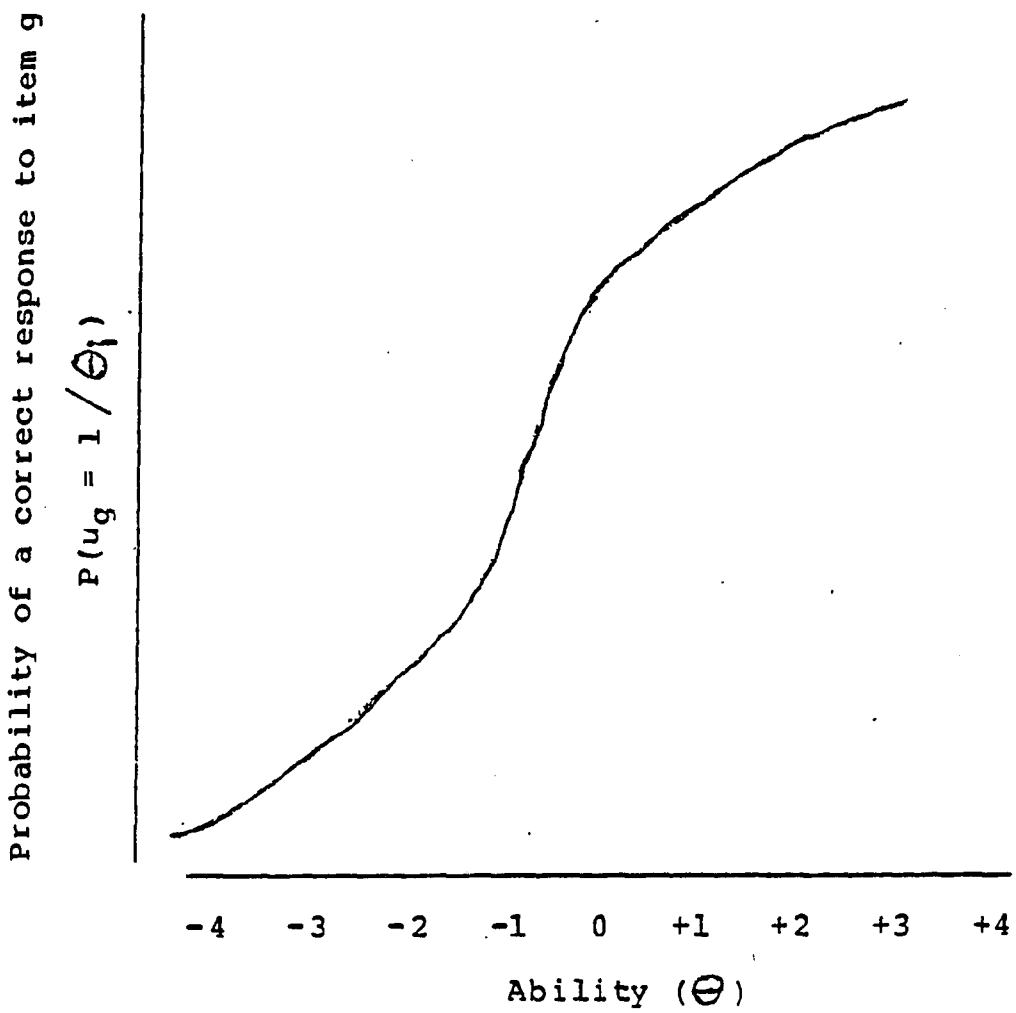


Figure 1. A Hypothetical Item Characteristic Curve (ICC)

underlying ability. Local independence refers to the assumption that an individual's response to an item on a test is independent of or uncorrelated with the individual's responses to other items on the test.

An item is considered to be unbiased if examinees of the same ability level, but of different group membership, have equal probabilities of providing a correct response to the item. In this case, the ICCs obtained from two different groups will be identical. If an item is biased, the ICCs for the two groups will not be the same.

Using latent trait theory, the probability of a correct response to an item may be defined using the three parameter cumulative logistic model proposed by Birnbaum (1968) and reviewed by Lord (1980). The formula for the logistic function relating item characteristics and ability to the conditional probability of passing an item is:

$$P(U_g/\Theta i) = c_g + \frac{1-c_g}{1 + e^{-1.7\, ag\, (\theta - b_g)}}$$

Where: a_g is the item discrimination parameter.

b_g is the item difficulty.

c_g is the item guessing parameter.

Parameter c_g, the guessing parameter, refers to chance level of response. It is the probability that a person completely lacking ability ($\Theta = -\infty$) will answer the item

correctly. Parameter b_g is the point of inflection on the curve or the location parameter. It refers to item difficulty and is the point on the ability scale where the probability of responding correctly to the item is .5. Parameter a_g is the discrimination parameter. It represents the discriminating power of the item and is proportional to the slope of the curve at the inflection point.

These parameters can be estimated by maximum likelihood procedures. As stated above, an item is unbiased if respondents from different groups, but the same ability, have equal probability of responding correctly to the item. Examples of item characteristic curves for items that are considered biased for two different groups are presented in Figure 2, 3 and 4. In Figure 2 and a_g parameters are unequal, that is the curves have different slopes. In Figure 3 there is an area between the curves because the b_g parameters are in different locations. In Figure 4, both the a_g and b_g parameters are unequal. The curves are not parallel and the inflection points are in different places on the ability continuum.

Various methods have been proposed for interpreting item bias using latent trait models. Rudner (1979) has suggested computing the area between the group item characteristic curves. He argues that a large area indicates that an item is not performing in the same way in the groups being compared. Lord (1977, 1980) has developed a

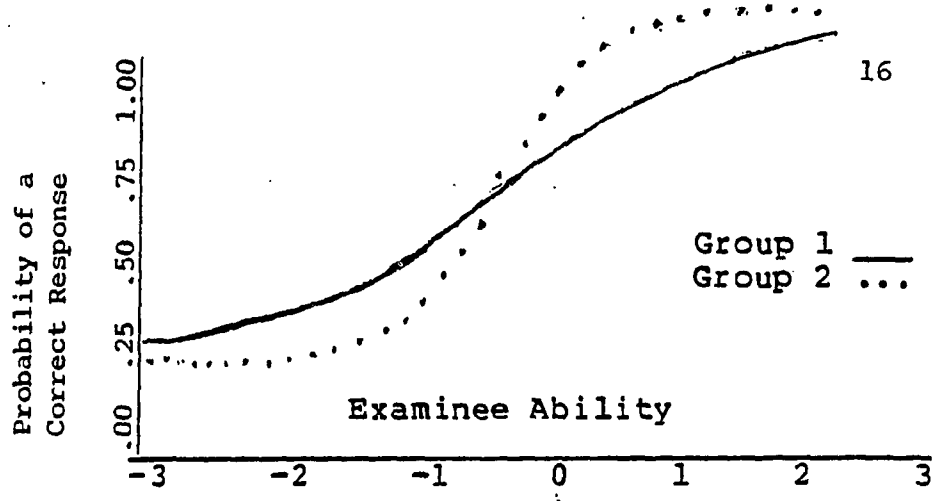


Figure 2. An ICC: Bias Solely in a_g

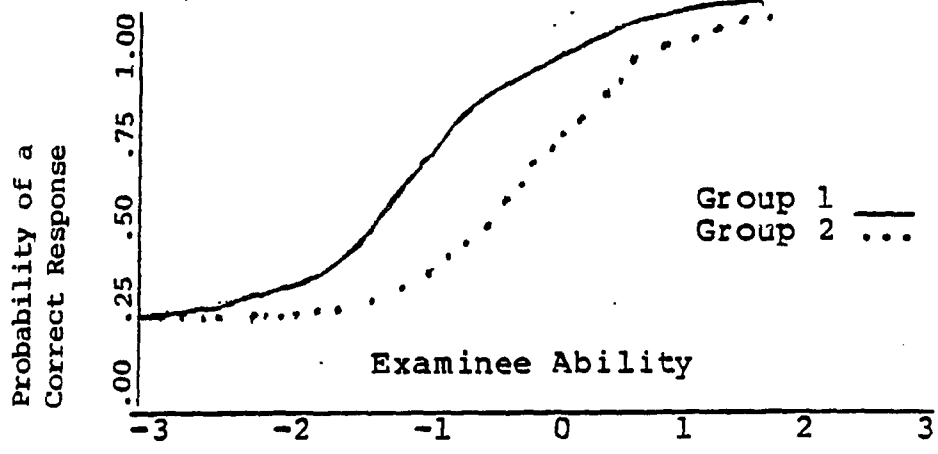


Figure 3. An ICC: Bias Solely in b_g

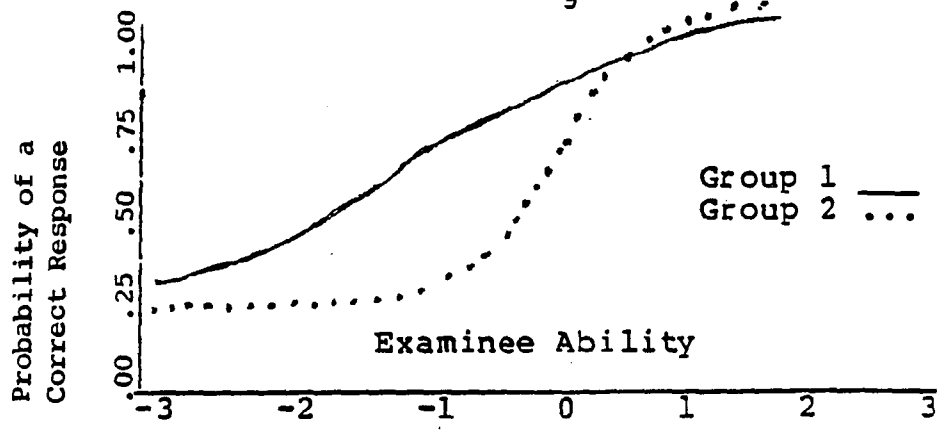


Figure 4. An ICC: Bias in Both a_g and b_g

(From: Merz, W.R., Grossen, N.E., Groome, M.L., and Groome, W.R. An Empirical Investigation of Six Methods of Examining Test Item Bias, 1978).

significance test, using a chi-square procedure, for determining the differences between ICCs. Using this method, differences in a_g's and b_g's for an item can be tested simultaneously.

Few studies have compared methods for identifying biased items. Nungster (1977) used data from the ninth grade state-wide test in Florida to compare several item bias identification models: Angoff's (1975) delta-plot procedures, Scheuneman's (1979) chi-square method, and Fishbein's (1975) p-value difference technique, where the difficulty values are compared. In attempting to identify items which contain sex bias, he concluded that all three models yielded similar results.

Rudner, Getson and Knight (1980) conducted a Monte Carlo study in which they compared three item characteristic curve approaches, two transformed item-difficulty methods and two chi-square procedures. A three parameter latent trait model was used to relate items and examined characteristics to item responses and to generate "bias." Conditions were varied in a number of ways. The correlations of generated "bias" and bias detected by each method were reported. Rudner et al. found that ICC model and the five-interval chi-square procedure were most accurate at detecting item bias across all conditions.

Angoff's delta-plot method, point biserial discrimination indices, Scheunman's chi-square model, and ICC tech-

niques were compared by Ironson and Subkoviak (1979). They used samples of black and white students who were tested as part of the National Longitudinal Study (1972). In general, there was little agreement between the methods in identifying biased items. The strongest agreement was between the three parameter ICC method and the chi-square model. With regard to the different models, they concluded that there are "no clear guidelines for choosing among them" and that practical considerations (e.g., cost) may best dictate the choice of model. While ICC models appeared to be the most accurate in detecting item bias, the use of latent trait (ICC) procedures may not be feasible for small scale studies because large samples are needed and the procedure is costly. Ironson and Subkoviack suggest that Angoff's delta-plot methodology and chi-square procedures are suitable alternatives in these situations.

Burrill (1980) compared several difficulty procedures (including Angoff's delta-plot, p-values, differences between delta-transforms), two correlation procedures (biserial and point-biserial) and Scheuneman's chi-square techniques for identifying biased items. She used black and white first-grade students and examined performance on the Metropolitan Readiness Test (1976 edition). Students were matched for ability on an external variable at the beginning of the investigation. Cotter and Berk (1981) have pointed out that a major concern with many item bias studies is

whether or not a significant item-by-group interaction is an indication of bias or an indication of initial ability differences between the subpopulations being studied. Angoff (1980) has also suggested the need for matching on ability in reviewing the work of Angoff and Ford (1973). By matching on ability, as Burrill has done, artifactual instances of bias can be reduced and sex or ethnic bias can be examined in an unconfounded form (Cotter and Berk, 1981).

Burrill found all of the difficulty procedures to be highly correlated with one another. The two discrimination indices also correlated highly, but the correlation between different kinds of approaches were not as high. Only chi-square correlated fairly well with each of the difficulty procedures. Burrill observed that different statistical procedures yield somewhat different information about which items are biased. It is important to note, however, that latent trait methods were not included in her investigation. Thus, using statistical item bias detection methods, the issue of item bias is considered in the absence of external criteria and is related to the properties of the items and their constancy across subgroups of interest. From a statistical point of view, bias may be thought of as invariance. Items which exhibit different properties across subgroups are considered to lack variance. An invariant item, on the other hand, refers to an item which behaves in the same manner and is inferred to be measuring the same underlying

ability or attribute across subgroups. Shepard (1980) stated:

> Statistical techniques for finding biased items are internal methods designed to ensure that the meaning which individual items contribute to the total test is the same for all groups . . .(However) the obligatory caveat for item bias methods is that they cannot detect pervasive bias because they lack an external criterion. . . . Item bias detection methods detect items that are anomalous. Whatever it is that the rest of the items measure, the biased item behaves differently.

In summary, there are four generally recognized statistical approaches to the examination of item invariance. From the simplest to the most complex, they are item discrimination indices, transformed item difficulties, chi-square procedures, and latent-trait theory or item-characteristic curves.

Bias detection methods based on item discrimination indices have involved the traditional correlation of item score and total test score--point-biserial correlations. A problem with this method is that it fails to take into account the ability level of the examinees. Thus, if two different groups have different levels of ability, one does not know if it is ability or some characteristic of the item which is causing the difference.

The transformed item difficulty-method most frequently used in detecting biased items is Angoff's delta-plot method. Using this method, item difficulties are calculated for two different groups. Confidence intervals are drawn around the major axis of the ellipse established for the two

groups. A decision about bias or invariance of an item is based on a preselected standard deviation cut-off. This approach has been criticized because there is no available statistical significance test for checking item bias.

Chi-square methods and item characteristic curve procedures differ from the above two methods in that they have the common assumption of an underlying trait or ability which is related in some way to performance on the items which comprise the measure. The response to an item is assumed to be reflecting some underlying ability. These methods differ from the above two in that they take ability level into account in assessing invariance. Chi-square procedures and item characteristic curve approaches differ from one another in terms of how they define ability.

Chi-square procedures divide ability into three, five or multiple intervals. Total score on a measure of latent ability is used to estimate ability intervals and a step function is used to relate ability to the probability of passing an item. Thus, an individual's total score on a test is considered, a decision is made about the individual's "ability" given the range of abilities, the individual is placed in an ability interval and the items for that individual are examined. Simply stated, according to the chi-square model, an invariant or unbiased item is an item where the probability of passing the item is the same for different groups at any given level of ability.

Item-characteristic curve approaches or latent trait theory methods for identifying biased items assume gradual increase in the odds of responding correctly to an item as an individual increases along the ability continuum. Thus, rather than a step function, as in chi-square, a continuous function is assumed. A curve representing the function that relates the likelihood of responding to an item to the underlying ability or attribute is known as an item characteristic curve. An item is considered to be invariant or unbiased if examinees of the same ability level, but from different subgroups of interest, have equal probabilities of responding correctly to the item. When this occurs, item characteristic curves are obtained from the subgroups that are identical or very similar. If an item is biased, the item characteristic curves for the subgroups will not be the same. Through visual inspection or through a statistical procedure using the item parameters, it can be determined how different or similar the curves are.

Judgmental Item Bias Methods

In addition to statistical methods for identifying biased items, investigators have begun to explore a priori judgmental methods for item bias identification. These investigators have contended that judgmental reviews are necessary to determine bias inherent in items which may be based in differential exposure of different gender and

cultural groups to material being tested. There is little research in this area and the evidence is conflicting as to whether judgmental and statistical methods for identifying biased items yield similar results.

Leinhardt and Seewald (1980) have studied the overlap between what is taught and test content. They have pointed out the necessity for a systematic investigation of the match between curricula and tests to determine if tests are assessing material to which students have been exposed and material with which they have had an adequate opportunity to become familiar. Test results cannot be meaningfully interpreted without this kind of analysis. The content validity of a test must be questioned when a test is measuring different material from that in which students are, or are to be, receiving instruction.

As study done by Bianchini (1977) suggested the importance of determining the differential content of tests. In this investigation, he examined the differential performance of first graders in California over a seven year period. During the first five years of the investigation, the Stanford Reading Test was used and the state-wide median for first-graders was at the 38th percentile rank on national norms. In the sixth year of the study, the Cooperative Primary Reading Test was substituted for the Stanford Test. The state-wide median for first graders was the same as that for the national norm sample. In exploring

the reason for this shift, Bianchini concluded that the norms on the Cooperative Test were not "easier" than those on the Stanford Test. However, the overlap between the Stanford Test and the first grade readers being used was 19 percent, whereas the overlap between the Cooperative Test and the readers was 55 percent.

Cooley and Leinhardt (1978) conducted a large scale study on reading and mathematics instruction in grades one and three. The results of their investigation indicated that the most meaningful explanation of achievement gains on tests in both subject areas was provided by the opportunity that students had had to learn the material assessed in the test. Those curriculum areas in which students received instruction were the areas in which test performance was adequate or good; students did poorly on test content in which they had never received classroom instruction.

A number of similar studies have been carried out at the Institute for Research and Teaching at Michigan State University. Porter et al. (1977, 1978) have developed an impressive taxonomy of elementary school mathematics curriculum and investigated and overlap between this curriculum and various standardized tests. They concluded that different tests measure very different aspects of the mathematics curriculum. In describing the material contained in the mathematics curriculum and on the mathematics tests, they use three factors to describe material: mode of presen-

tation, nature of material, and operations. Interrater reliability was established and all ratings involved classification by at least three researchers. Tests themselves were found to be reasonably consistent with one another in their coverage of whole numbers and computations skills, but different in their coverage on the use of graphs, tables and numbers sentences.

Freeman et al. (1980) developed a similar classification system for examining and coding the overlap between tests and commonly used texts. In addition, materials have been developed to identify the relevant characteristics of classroom curriculum (other than text book materials) and to describe the overlap of this curriculum and tests (Kuhs et al., 1979).

McCarthy (1975) has shown that the performance of high school males and females on mathematics problem-solving tests is influenced by the content of the test items. This is caused by differential familiarity of subgroups with the context in which the mathematics problem is presented and by differential exposure to the subject area.

Tittle (1975, 1980) also addressed the overlap between instructional materials and test items. She pointed out that decisions about the match between curriculum and test material, as well as decisions about amount of exposure that students have had to material, require judgmental methods. In that the use of expert judgment has always been applied

in test construction (to item tryout forms and to the selection of items for final forms), this approach is not new. However, the systematic use of these judgments throughout the test development and revision process, given the problems about validity outlined above, is an important innovation.

Tittle has proposed that panels of judges representing the different gender, ethnic, and socioeconomic groups to be tested should review each of four areas:

1. To rate items and tests for equal exposure of subgroups to whom the test will be administered to have learned material covered on the test.

2. To rate items and tests for equal opportunity for all subgroups to whom the test will be administered to have learned material covered on the test.

This examination is needed to establish the content and construct validity of the instrument. In addition, these panels should review tests for two additional purposes:

3. To rate items and tests for stereotyping of subgroups.

4. To rate items and tests for fair representation of subgroups.

This review should contribute to establishing the face validity of the instrument. Judges should be given explicit instructions about how to rate each item using item review

scales developed for this purpose. The review of a test and/or test items will necessitate an ongoing process of negotiation and accommodation on the part of the judges at the various stages at which ratings are applied. Tittle suggested that these stages should include: test planning, item writing, item review, item selection for final forms, norm development and the reporting of test results.

In addition, panels may be formed to review items and the test (at all stages of test development) for content and construct validity. These panels should be made up of curriculum experts and other individuals familiar with the material to which students have been exposed in the classroom and knowledgeable about the areas in which students have received instruction.

Judges should not only be used in reviewing items, but at other decision points. In the item tryout stages, a crucial decision must be made as to which major subgroups will be involved in the piloting of the test. In addition, a judgment must be made as to which statistical procedures to use in analyzing the test results for possible bias (see the section on statistical item bias methods above). In the development of norms, as with the piloting of the test, judgments determine the choice of subgroups to be included. Tittle has suggested that judges should be involved in each of these decisions.

Cole (1980) has also emphasized the need for expert judgments in the test construction process. She has stressed (1) the need for training panel members to be sensitive to item bias concerns and (2) the need to document all efforts taken to eliminate bias in the test development process, particularly the comments and reactions of judges. Cole has pointed out the necessity of linking the constructs of familiarity and opportunity to learn with statistical notions of fairness.

In order for this to occur, judgmental and statistical data must be compared. However, little research has been done on the systematic use of ratings throughout the test development and revision process and only recently have studies been undertaken to compare judgmental methods with statistical methods of bias identification. These studies will be reviewed in the next section.

New Approaches to Identifying Biased Items

Test constructors and publishers have been concerned with developing tests that are unbiased both from a facial review and a statistical analysis. Thus, much of the pioneering work that has occurred concerning the similarities and differences between perceptions of bias and statistical indications of bias has been done by researchers involved in the development of standardized tests.

Schmeiser (1980) has conceptualized a three-stage approach to the identification and elimination of biased items in tests. She labeled the judgmental approaches to the elimination of bias as "a priori" item bias procedures because they are geared toward detecting bias during the early stages of test development. She labeled the statistical approaches in the elimination of bias "ex post facto" item bias procedures because they occur after the test items have been administered. Schmeiser has pointed out the need for an intermediate stage in the elimination of item bias. Since this method is based on principle of experimental design, she has referred to it as "the experimental design" approach. This intermediate stage should involve the systematic investigation of procedures used by test developers in the construction of tests and should identify any changes that are needed in the test development process.

Schmeiser has explicated the experimental design procedure as it could/should be used in test development efforts. Schmeiser's "experimental design" approach to test development does not provide a new or different kind of procedure for identifying the eliminating biased items, but it involves a comprehensive discussion of the necessity for merging judgmental and statistical methods for identifying biased items in test construction efforts. In addition, she makes an important point in suggesting that information learned in the development of one test should be used to

modify and improve the procedures through which all tests are constructed.

Angoff (1980) has strongly urged that judgments and statistical methods for r identifying biased items be combined. He states that "statistical methods are only supplemental to human judgment, certainly no substitute for it." He has argued that proper test development requires that all items be reviewed by judges at several stages for ethnic and sex bias before they are included in a test. Angoff has suggested that outlier items be reviewed to see what kinds of skills they are tapping:

> There should be an educational or psychological rational for deciding that a statistically biased item is indeed biased. Consider the fact that in some of our analyses it has been found that items having to do with percentage appear to be especially more difficult for blacks than for whites, relative to other items in tests. Guided by statistics one might conclude that since items are "biased" they should be removed from the test. On the other hand, if these items are tapping knowledge and skills that are needed for survival in our society, then, biased in the statistical sense or not, they should remain. It is plain that the problem revealed by the statistics in the analyses cited here lies not in the items but in the quality of the training that students receive in dealing with knowledge and skills (p.27).

The major test publishers have responded to the test bias issue by increasing the scrutiny to which items are subjected during the test construction process. Editors and reviewers are being trained to be sensitive to issues of stereotyping and to material that is potentially offensive to different subgroups. Educational Testing Service, for

example, has developed guidelines which have been distributed to all of their staff members governing these areas as part of their sensitivity review process (Fremer, 1980; Hunter and Slaughter, 1980). McGraw Hill has developed similar guidelines for its staff, including the Guidelines for Equal Treatment of the Sexes and the Multiethnic Publishing Guidelines. In addition, test publishers have tightened statistical review procedures for identifying biased items and, recently, they have begun studies comparing various methods of identifying item bias.

Green (1980) has reviewed the procedures used by CTB/McGraw Hill for identifying and eliminating item bias. He has provided a detailed description of the item review procedures used by CTB/McGraw Hill and has reported on the statistical results of item tryout studies comparing responses of members of different ethnic groups. To date, these studies have primarily involved the comparison of results of black and white test-takers. Green has pointed out the difficulty CTB/McGraw Hill has had in identifying and gaining access to students of other ethnic groups. He has suggested the need for additional research involving these groups. At CTB, statistical analyses have been performed using both chi-square and latent trait (three parameter ICC) models.

Coffman (1980) discussed the research that has been undertaken by the Riverside Publishing Company at the

University of Iowa to develop less biased assessment instruments. Researchers have begun to compare judgments of representative panels of test users as to the fairness of items on tests with statistical indicators of item bias. He pointed out the difficulty that item writers are having in developing "fair" items, given the many regional differences that exist in language, culture, and school curricula throughout the nation.

The Riverside Publishing Company has recently experimented with a review panel which consisted of the following ethnic breakdown: nine black reviewers, three Native American reviewers and three Asian reviewers. Half of the reviewers were female, half of the reviewers were male. Each reviewer was asked: "to make judgments from a broad frame of reference, not simply from the viewpoint of a particular racial background" (Coffman, 1980). Data collected from the reviewers served to generate hypotheses for further investigations of the source of item bias. He concluded that there is a great deal more to learn from the input of panels of experts with varied cultural backgrounds.

Several problems with the Riverside Publishing Company use of a single panel of judges with diverse ethnic backgrounds are immediately apparent. First, judges were given directions to be sensitive to a wide variety of backgrounds, but they were given no training as to how to do this.

Tittle (1980) and Cole (1980) have pointed out the necessity of training, if judges are to be attuned to potential bias. Second, it is not clear that it is reasonable to expect a panel of individuals with diverse ethnic backgrounds to be sensitive to or aware of material that may be offensive to members of other ethnic or racial backgrounds. It would make more sense, as Tittle and others have suggested, to make use of several panels for this purpose. Each panel would consist of members of a particular subgroup of the population to which the test will be administered. Lastly, an examination of a mathematics test published in 1979 by the Riverside Publishing Company revealed that while the use of names showed multiethnic and non-sexist awareness, the language on the test was confusing and regional. Local names which would not be known in New York City, for instance, were used for common foods, including the term "drumsticks" for ice cream cones and "poor-boys" for hero sandwiches.

The investigations undertaken to explore the use of judgmental approaches in conjunction with statistical approaches in test development and revision have been inconclusive. The importance of judgmental methods to supplement statistical methods has been established, but the evidence on how and when to employ the ratings of experts remains sketchy and unresolving.

Qualls and Hoover (1981) have explored the use of ratings of test items by black and white teachers for potential race favoritism. Using the Iowa Test of Basic Skills (ITBS), they asked 41 white and 51 black teachers to rate each item on the test for potential race favoritism. They used a five point scale which ranged for "definitely favors whites" to "definitely favors blacks." Qualls and Hoover (1981) investigated whether teachers' ratings interact with teacher race, item types, and subtests. They found that the results from the six subtest ANOVA's identified race by item-type interaction for the spelling and capitalization subtests. On both subtests, black teachers indicated "larger degrees of favoritism, which could suggest a greater sensitivity toward cultural differences." Qualls and Hoover concluded that more precise and systematic investigation is needed of role of the reviewers in bias identification.

Alterman and Holland (1981) examined the Tests of English As a Foreign Language (TOEFL) for instances in which item performance of examinees with comparable scores differed according to their native languages. They employed a chi-square technique and the use of the ratings of reviewers familiar with particular languages. They found that reviewers familiar with specific languages were able to find reasons for item performance differences post hoc, but reviewers were not successful in identifying items where differential performance might occur without having the

results of the test. Given the unique nature of this test, and the six very different language groups involved, these findings are not surprising. Alderman and Holland (1981), in pointing out the complexity of the reviewers' task, suggested that, ". . . explanations necessarily depended on each reviewer's familiarity with the several native languages and the linguistic similarities and dissimilarities with the sound, syntax, semantics, and vocabulary of the English language" (p. 28). They concluded that detecting instances of, and determining reasons for, differential item performance of different language groups requires the use of an appropriate statistical technique as well as meaningful rating procedures. However, in this study it appears that the task of the judges was especially difficult because of the multidimensionality of language transition.

Cotter and Berk (1981) investigated item bias in the Wechsler Intelligence Scale for Children (1974) (WISC). They used black, white and Hispanic, learning-disabled, ten-year-old students, matched on subtest raw score and sex. An item-by-item group analysis of variance was computed and ratings of educators from the three subgroups were collected. The black and white committees met as a group to discuss each item separately. The Hispanic committee responded individually and only to those items that were statistically identified as biased.

Cotter and Berk (1981), in integrating judgmental and statistical methods, found that bias existed in four subtests in the black-white comparison and three subtests in the Hispanic-white comparison. A shortcoming of this investigation was the lack of formalization of the rating process. Two of the subgroups worked as committees and reached their decision by consensus; the third subgroup submitted individual ratings on those items identified as biased through a statistical procedure. In order to collect data that can be meaningfully compared between the different subgroups, the use of educators' judgments should have been systematized and consistent throughout the investigation.

Statement of the Problem

A review of the above literature in the areas of item bias suggests the need for research in which the statistical and judgmental methods for assessing the fairness of tests are compared. While until recently, only one method or the other was used to identify biased items, researchers and the major test publishers have begun to employ both procedures. However, these procedures have not been employed systematically. Research is needed to determine if the same or different items are identified as biased using each method; to determine how the elimination of biased items affects the test scores of different subgroups of the population; and to make explicit the purposes for which statistical and judgmental procedures should be employed.

In order to meet this need, the proposed study will be undertaken to compare items identified as biased on a criterion-referenced mathematics test using statistical and judgmental methods. The results of a statistical item analysis of a test, with regard to differential performance of different subgroups, will be compared to the ratings of expert judges of the fairness of each item on the test for these subgroups. The test will be revised based on these data. The statistical analysis will be conducted using

latent trait methods for item bias detection because the results of previous investigations suggest that this is the preferred method for large samples. For the judgmental analysis, rating scales to identify perceived item bias or lack of content match will be employed formally throughout the investigation.

In summary, the present study is proposed:

1. To compare judgmental methods and statistical methods for identifying biased items.

2. To combine these methods in revising an existing criterion-referenced test.

3. To determine if a test that is rescored to eliminate items identified as biased for different subgroups of the population will yield equal average scores across the subgroups.

It is only when the relationship between statistical and judgmental methods is understood that they can be meaningfully used to detect item bias in the test construction and revision process. The elimination of bias is essential to the fair use of tests in providing equal opportunities to different subgroups of the population.

Critical Questions to be Investigated

1. Will the test items identified as biased by reviewers from different subgroups of the population be related across subgroups?

2. Will the test items identified as biased by a statistical procedure for ethnic subgroups of students be related to those identified as biased for gender subgroups of students?

3. Will statistical and judgmental methods for identifying biased items on a test agree?

4. Will a test that is rescored eliminating items identified as biased yield significantly higher student outcome results for each subgroup and for the total group?

5. Will student performance on a criterion-referenced test in mathematics and on a rescored version of the test, eliminating items identified as biased, be correlated with student performance on the mathematics subtest of the Stanford Test of Academic Skills, Level 1, (1971) and will these two versions of the test be correlated with one another?

Method

Subjects

The sample was composed of 1,064 male and female high school students, drawn from grades 9-12 in fifty remedial mathematics classes. The students were selected from New York City high schools which serve remedial students. The students were of diverse ethnic backgrounds (e.g., black, Hispanic, other) and were two or more years below grade level in mathematics as determined by a standardized norm-referenced mathematics test.

Procedures

Overview. In Step I, the New York City Criterion-Referenced High School Mathematics Test (Test 1) was revised by the staff of the remedial mathematics program based on existing data, so that the new test (Test 2) better matched the high school mathematics program curriculum. In Step II, Test 2 was administered to 1,064 high school students of both sexes and different ethnic groups who participated in the remedial mathematics program. The results of the administration of Test 2 were analyzed to determine the psychometric properties of the test. For each subgroup, the item characteristic curve was compared for each item. Test 2 was examined by three groups of reviewers: blacks,

Hispanics, and whites. Half of the members of each group were male, half were female. They rated each item on the test for content familiarity and ethnic or gender bias.

In Step II, the test was rescored by subgroup, eliminating those items identified as biased, to determine if differences occurred in score distribution.

In addition, the results of statistical and judgmental methods of bias item identification were analyzed and compared. These procedures are presented in Figure 5 and discussed more fully below.

Step I: Background - Content Analysis and Revision of Test I - Judgmental Review and Statistical Review. The New York City Criterion-Referenced High School Mathematics Test (Test 1) was reviewed by a group of eight mathematics content experts who are staff development specialists in the New York City remedial mathematics program serving all high schools in New York City and by the coordinator of the program. As a group, they reviewed each item on the test to determine if the items were over-representing or under-representing areas of the program curriculum, as well as to judge the appropriateness of the difficulty of each test item. The group made decisions about items to retain, to delete or to rewrite based on the match to program curriculum and the judged appropriateness of the difficulty level of the item.

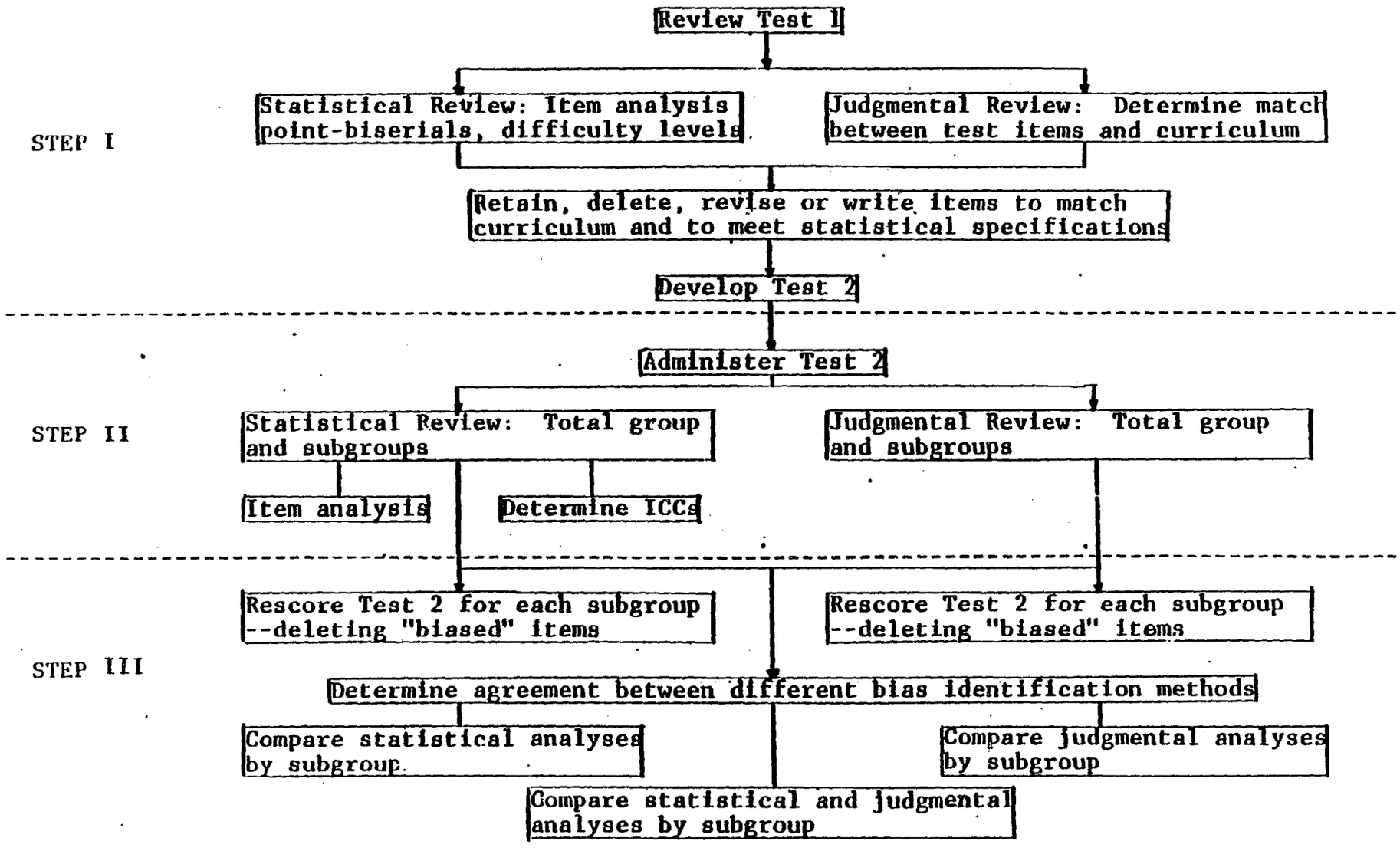


Figure 5. Experimental Procedure

An item analysis was performed on Test 1 using extant data on 300 students according to an existing New York City Board of Education procedure used when the test was developed. The item difficulty and item discriminability of each item were computed. An item was retained and included on the revised test (Test 2) if the item was judged to be content appropriate and if the item had a moderate item-total correlation, ranging from .22 to .60, and a mean passing score of about .50. An item was deleted if it was not psychometrically sound or did not match the curriculum. An item was revised if: a) it was not statistically a good item, but it matched the curriculum and the psychometric problems with the item could be tentatively identified (e.g., poor distractors, unfamiliar format) or b) it was statistically sound, but did not match the curriculum and could be brought into accord with the curriculum with little substantive change.

Based on the review of Test 1 described above, fifty-eight of the original 129 items were maintained as they had existed on the original test. These items met the stipulated statistical criteria and were judged by the content experts to be appropriate. Seventeen items from Test 1 were revised so that they better matched the program curriculum or in an effort to improve their psychometric properties. Fifty new items were written: these items were written to replace items that were statistically poor items or were not

content appropriate items and to tap areas of instruction that were not being adequately assessed on Test 1.

The new test, Test 2, developed by the staff of the remedial mathematics program in consultation with other offices of the Board of Education contained 124 items. However, a modification took place in the middle of the data collection for this investigation. Based solely on content considerations, the program staff changed three items at the end of the test and added an additional item, creating four new items. This modification provided 121 common items given to all subjects and available for analysis by item response techniques in this study. Judgmental ratings for all 125 items from the latest field version of the test were elicited from reviewers in the interest of having as broad a range of items as possible when evaluating interjudge reliability. Of course, in comparing the statistical and judgmental procedures for identifying item bias, only the 121 common items were used.

Step II: Administer and Review Test 2

The present investigation began with the administration of and a formal review of Test 2 (see Appendix I). The procedures followed are described below.

Judgmental Review. Test 2 was reviewed by twenty-four individuals who are knowledgeable about high school mathematical curriculum for material that may be too difficult

for or unfamiliar to high school students in basic skills classes and for material that seemed offensive and/or stereotypical. Eight of the reviewers were black, eight of the reviewers were white, and eight of the reviewers were Hispanic. Within each of the groups, half of the reviewers were male and half were female.

Reviewers in all groups had a broad knowledge of high school mathematics curriculum in New York City. Most of them were responsible for supervising or teaching the implementation of the mathematics curriculum, including college instructors of prospective mathematics teachers, assistant principals for mathematics, mathematics department chairpersons, teacher trainers and master teachers in mathematics.

Each of the judges was provided with a review package containing guidelines for identifying ethnically or sexually offensive material adapted from ETS Sensitivity Review Procedures and a structured rating form for judging items that may be unfamiliar or offensive to students in basic skills classes.

The rating form employed for the judgmental review procedure was adapted from a scale developed by Hambleton (1980). The instrument was refined in order to clearly identify different kinds of perceived bias (gender, ethnic, curricula) and in order to attempt to distinguish item bias

from item difficulty. Hambleton's procedure did not provide for this distinction.

The new rating form was piloted on six mathematics teachers. The teachers were asked to review a subset of items from the Advanced Arithmetic Test of the Metropolitan Achievement Tests (1958). There were fifteen computational items and fifteen word problems included for review. Data from the pilot indicated that reviewers, using the modified review form, identified different kinds of perceived bias (gender, ethnic, curricula).

For computation items, judges were asked to respond "Yes" or "No" to three questions for each item:

1. For high school students in basic skills classes is this an especially difficult item?

2. For high school students in basic skills classes is item content likely to be new to a particular subgroup?

3. For high school students in basic skills classes is the item format likely to be new to a particular subgroup?

If judges responded "Yes" to any of the questions pertaining to an item, they were asked to explain in writing the reason for their judgment.

For word problems, judges were asked to respond "Yes" or "No" to eight questions about each item:

1. For high school students in basic skills classes is this an especially difficult item?

2. For high school students in basic skills classes does the item contain difficult vocabulary and/or sentence structure?

3. For high school students in basic skills classes is the item content likely to be new to a particular subgroup?

4. For high school students in basic skills classes is the item format likely to be new to a particular subgroup?

5. For high school students in basic skills classes does the item contain language which would be offensive to a particular subgroup?

6. For high school students in basic skills classes does the item contain descriptions which would be offensive to a particular subgroup?

7. For high school students in basic skills classes does item contain activities or situations which would be offensive to a particular subgroup?

8. For high school students in basic skills classes does item contain language which would be new or lack common meaning to a particular subgroup?

Again, if judges responded "Yes" to any of the questions pertaining to an item, they were asked to explain in writing the reason for their judgment.

The ratings of the judges to each question for each item were combined by ethnic subgroups, by gender subgroups, and for the total group.

Statistical Review

Test 2 was administered to 1,064 high school students of both sexes and diverse ethnic backgrounds enrolled in New York City public schools. Complete data were obtained from 956 students. This group consisted of 463 males, 493 females, 521 blacks, and 435 non-blacks. These results were to be analyzed to estimate item and ability parameters using the LOGIST program developed by Wood, Wingersky and Lord (1976). LOGIST is the program used to obtain parameter estimates by maximum likelihood procedures when the three-parameter logistic model is employed. As discussed above, the formula relating item characteristics and ability to the conditional probability of passing an item is:

$$P(U_g/\theta_i) = c_g + \frac{1 - c_g}{1 + e^{-1.7a_g(\theta - b_g)}}$$

Where: a_g is the item discrimination parameter

b_g is the item difficulty parameter

c_g is the item guessing parameter

A number of significance tests were then to be performed to test and hypotheses that the a_g's and b_g's for a particular subgroup were equal to the a_g's and b_g's for the rest of the sample tested; that is the total group excluding the sub-group being compared (T·j).

However, the latent trait theory models are based on the assumption of unidimensionality (Lord and Novick, 1968). The test that was employed in the current study was a criterion referenced test. It did not contain one factor, but was multidimensional. This became apparent when the LOGIST procedure was used and the model fit was tested by comparing the a_g's and b_g's between arbitrary subsets of subjects. Many more chi-squares were significant than would be expected under the null hypothesis given the unidimensional model. Kife and Brumble (1974) have suggested that multidimensionality is frequently cited as one of the reasons why data does not fit the model. Reckase (1977) documented this point in an extensive study in which he used both simulated and real data with different factorial structures.

Specifically, in the current study when the LOGIST model was run on the test data the a_g's were much larger than those usually obtained according to Lord and Wingersky (personal communication). On an odd/even split of subjects, forty-one significant chi-squares were found out of 121 items ($\alpha = .05$). The data were not fitting the model due to the multidimensionality of the test.

Several approaches were explored in attempting to handle the test's multidimensionality. First, the test which is composed of three subsections (whole numbers, fractions, and decimals and percentages) was separated and the procedure was run independently for each subsection. How-

ever, instead of reducing the difficulties due to multi-dimensionality, this increased the problem; reducing the number of items made the remaining minor factors relatively more important and resulted in a comparably high proportion of significant chi-squares among the items in a subsection. Thus, this procedure was dropped.

A second attempt was made to respond to the lack of fit of the data to the model caused by the multidimensionality of the test following suggestions proposed by Wingersky (personal communication) and Reckase (1977). Instead of allowing both the a_g's and the b_g's to vary between subgroups, the a_g's were estimated for the whole sample and the b_g's were allowed to vary between groups. The procedure should be superior to using the simple logistic model (the Rasch model) because it doesn't require that all the items are equally discriminatory. Furthermore, because the items on the test were multiple choice items and the simple logistic model does not contain a guessing parameter, the use of the simple logistic model was undesirable.

The analysis was then run holding the a_g's constant and the c_g's constant for an odd/even split of subjects. The a_g for each item was held at the value determined for the whole group. There were, however, six items of the 121 items on the test that were problematic; five of these items had poorly determined c_g's (larger than would be expected from random guessing) and one item had an outlying high value for

a_g. Therefore, good estimates of the b_g's for these items could not be expected. That is, the rate at which a person of zero ability would get these items correct was better than that of guessing, so estimates of the difficulty parameters for these items could be expected to be biased. For this reason and because the precision of the estimates of the other items might be reduced by these items in that all parameters are solved simultaneously, a decision was made to remove these items from further analysis.

Following this, the program was run on an add/even split of subjects. This resulted in a number of significant z's ($\alpha = .05$) near the value expected by chance. The program was then run for black and non-black groups of subjects. There were three times as many significant z's as for the odd/even split of subjects. An analysis of the male and female groups was performed next. This analysis resulted in twice as many significant z's as the odd/even split of subjects.

The relatively small number of Hispanic students (N = 215) who took the test did not allow for an analysis of the Hispanic subgroup that would yield stable results using maximum likelihood procedures which is a large sample technique. Furthermore, the comparison of the Hispanic group to a reference group of non-Hispanic students potentially entailed an additional problem: this comparison would have influenced the comparison of the black group to its refer-

ence group. That is, in order for the black and other-than-black comparison not to be seriously confounded with the Hispanic and other-than-Hispanic comparison, the other group in both of these comparisons would have to be (non-Hispanic) whites. This would have reduced the size of the comparison group for the blacks, as well as for the Hispanics. Although this would have been the most correct form of the analysis, the stability of the results of the black and other-than-black analysis would have been impaired by the reduction in size of the reference group. Therefore, it was decided to make one relatively stable comparison, rather than two unstable, confounded comparisons.

Thus, a method of analysis which provided a reasonable baseline for estimating how many items would be identified as biased due to sampling fluctuation or lack of model fit was established. This model further identified a number of items in excess of the base rate as being biased in the black and non-black groups of subjects and in the male and female groups of subjects.

In summary, the final method of analysis held the a_g for each item at the value determined for the whole group while the b_g's were allowed to vary between groups. Six items were removed from the analysis; five because they had poorly determined c_g's and one because it had an outlying value of a_g.

Step III: Data Analysis

The characteristics of individuals that are commonly examined in the item bias work are gender and ethnicity. In this study, the gender groups were male and female and the ethnic groups were black, Hispanic, or other. The question of item bias was addressed, using item response theory techniques, by examining the behavior of items across subgroups which differ on one characteristic and contained students who were heterogeneous on the other characteristic (e.g., gender). Similarly, judges ratings were analyzed for subgroups of judges defined by one characteristic (e.g., gender) who were heterogeneous on the other.

Ratings for judges were combined in a number of different ways in order to compare agreement in the identification of bias between judges within groups, across groups, and with statistical results for different subgroups. The first indicator was chosen to use all the information about bias that was available uniformly across all items. As with all indicators, it was defined so that it was not idiosyncratic to a particular rater, but represented bias as detected by raters who may be presumed to have similar concerns. Within groups of raters who were homogeneous with respect to one characteristic (i.e., ethni- city or gender), an item was considered to be biased if at least two raters within a homogeneous group considered it biased. (In this case, the indicator was assigned a value of one, as opposed to zero.)

The reviewers, on the whole, rated relatively few of the questions as biased. Three of the twenty-four reviewers assigned no bias ratings to any of the items on the test. Most of the reviewers responded "No" to the questions about computational problems, with an occasional "Yes" response given to one of the questions about a particular item. The questions about word problem items elicited more "Yes" responses from the reviewers. The frequency of items judged bias by the reviewers is discussed below. Given the overall response pattern of the reviewers, a liberal decision rule was used to combine the reviewers' ratings in computing the bias indicators to ensure that all available information was used in being sensitive to bias concerns. The rule was designed to be sensitive across subgroups with regard to perceived bias.

The first indicator (I_3) was based on the three questions judges were given about each computational item and the corresponding questions about word problems (questions 1, 3, and 4). These questions were:

- Is this an especially <u>difficult</u> item?
- Is the item <u>content</u> likely to be <u>new</u> to a particular subgroup?
- Is the item <u>format</u> likely to be <u>new</u> to a particular subgroup?

If two or more judges within a particular subgroup responded "Yes" to any of these questions about an item, the item was

considered to be biased for that subgroup ($I_3 = 1$, otherwise $I_3 = 0$). Then an overall index was defined with a value of one if any of the subgroups had an index of one, otherwise the overall index was zero. This procedure was applied to subgroups based on ethnicity (blacks, Hispanics, others) generating one overall index and to subgroups based on gender (males and females) generating a second overall index. The overall index based on ethnic subgroups guaranteed sensitivity to concerns varying across ethnicity and, similarly, the overall index based on gender subgroups guaranteed sensitivity to concerns varying across gender.

Following this same set of rules, additional within group and overall indicators (I_8) were computed for the word problems. For these indicators, all eight questions posed to each judge in rating the word problems were used, as opposed to just the three questions employed computing I_3. Thus, if two or more judges within a subgroup responded "Yes" to any of the eight questions about a word problem item, the item was considered to be biased for that subgroup ($I_8 = 1$, otherwise $I_8 = 0$). Similarly, this procedure was repeated twice, once for ethnicity and once for gender in generating overall indices.

A simpler set of indicators (Q_1) were also computed for each item using only the first question to which judges were asked to respond about each computational item and word

problem item:

Is this an especially difficult item?

It was believed that responses to this single question were most likely to be tapping similar information to that identified by the statistical procedure; that is to be sensitive to items that were especially more difficult than other items on the test. Again, if two or more judges within a subgroup responded "Yes" to this question about a given item, the item was considered to be biased. This procedure was also repeated twice, once for ethnicity and once for gender.

The frequency with which items were judged biased by subgroups (male, female, black, Hispanic, other) for indicators Q_1, I_3, and I_8 are presented in Table 1. For each indicator, an item was considered to be biased by a given judge if the judge responded "Yes" to one or more questions about an item. Each ethnic subgroup was comprised of eight judges, so the frequencies of bias ratings for an item could range from 0 - 8. Gender subgroups were comprised of twelve judges, so the frequencies of bias ratings could range from 0 - 12. Indicator Q_1 was based on only one question, indicator I_3 was based on three questions, and indicator I_8 was based on eight questions. Thus, it was most likely that I_8

Table 1

Frequency With Which Items Were Judged by Subgroups for Indicators Q_1, I_3, and I_8

Frequency With Which Items Were Judged Biased	Q_1					I_3					I_8				
	M	F	B	H	O	M	F	B	H	O	M	F	B	H	O
0	108	83	99	99	102	56	69	69	71	81	8	7	5	8	8
1	12	23	23	24	18	55	30	29	37	26	10	4	12	6	11
2	4	17	3	2	5	6	18	6	16	11	8	2	8	10	8
3	1					7	5	1	1	7		6	4	2	
4		2				1	3				2	5	1	1	1
5												2		1	2
6											1	2		2	
7											1	1			
8												1			
9															
10															
11															
12															
Total Items	125	125	125	125	125	125	125	125	125	125	30	30	30	30	30

Legend
M = Males
F = Females
B = Blacks
H = Hispanics
O = Others

would indicate bias in that if one of eight questions about an item received a "Yes" response from a given judge, the item was considered biased. The data presented in Table 1 demonstrate the low frequency with which items were considered to be biased by the judges who reviewed the test. For example, 108 of the 125 items were never judged biased by any male judges on indicator Q_1; 12 items received one "bias" judgment from the male judges on this indicator; two items received four "bias" judgments; and three items received one "bias" judgment. This pattern of infrequent ratings of items as biased is repeated across subgroups and across indicators.

Agreement between statistical bias ratings among subgroups, agreement between judgmental ratings among subgroups, and agreement between item bias identification by subgroup characteristics (i.e., ethnicity or gender) using the statistical procedure and the judgmental procedure were assessed using the Kappa statistic (Cohen, 1960) or the intraclass correlation coefficient. Under a wide number of conditions, Kappa is equivalent to the intraclass correlation coefficient. The case in which ratings are dichotomous is a sufficient condition for this equivalence (Fleiss and Cohen, 1973; Davies and Fleiss, 1981).

Kappa is the proportion of agreement between different raters or procedures corrected for chance. Kappa varies

between negative values and +1; +1 indicates perfect agreement, zero indicates exactly chance agreement, and negative values indicate less than chance agreement. Kappa may be defined as:

$$K = \frac{P_o - P_c}{1 - P_c}$$

Where: P_o is the proportion of observed agreement

P_c is the proportion of agreement expected by chance.

Using the simplest example, if two judges (J_1 and J_2) rate the items on a test as biased (B) or unbiased (U), the agreement between the ratings may be represented as:

		J_1: B	J_1: U		
J_2	B	P_{11}	P_{12}	$P_{1.}$	$P_o = P_{11} + P_{12}$
	U	P_{21}	P_{21}	$P_{2.}$	$P_c = P_{1.}\ P_{.1} + P_{2.}\ P_{.2}$
		$P_{.1}$	$P_{.2}$		

The agreement between the ratings of two groups of judges (e.g., males and females) may be represented in the same way. The ratings of individual judges were collapsed according to the rule discussed above to compute within subgroup indicators. An item was defined as biased in a particular subgroup for any of the three indicators if two or more judges within the subgroup responded "Yes" to a question about the item. Thus, the cells in the above table

contain the number of items with the four possible combinations of ratings by male and female subgroups (e.g., judged biased by both men and women, judged biased by men and unbiased by women, etc.).

Where there are three groups of judges (e.g., black, Hispanic, other), the observed and expected agreements are defined in terms of the average number of agreements for all possible pairs of groups. Specifically, in this case there are three possible pairs to be considered: black-Hispanic, black-other, and Hispanic-other. For a dichotomous judgment, the case of three groups may be represented by a table of X_{ijc} with cell entries in row i and column j for one of c categories. If c = 1, meaning biased, the table is:

	Biased			
	B	H	O	Y_i
Item 1	0	1	0	1
	0	0	0	0
	1	0	0	1
	0	1	1	2
	1	1	1	3
125				
	P·1	P·2	P·3	

The observed and expected agreement (e.g., P_o and P_e) between the ratings may be represented as the average value of the three entries in the lower triangle of a 3 x 3

matrix:

	B	H	O
B			
H	P21		
O	P31	,P23	

For example, $P_{c21} = P_{\cdot 1} \cdot P_{\cdot 2}$ and

$$P_c = \frac{P_{c21} + P_{c31} + P_{c23}}{3}$$

The formula for treating the more general cases of non-dichotomous judgments and more than two sources of ratings was developed by Davies and Fleiss (1982). It does not look like the simpler formula for Kappa presented above because it is derived from more complicated formulas, but it can be shown to be algebraically equivalent to the simpler formula. The formula for Kappa developed by Davies and Fleiss is:

$$K = 1 - \frac{Y_{ic}(J-Y_{ic})}{I\left[J(J-1)\,\bar{P}_c\,(1-\bar{P}_c) + \sum_j (P_{jc} - \bar{P}_c)^2\right]}$$

Where: I = number of items

c = number of categories

J = number of groups

X_{ijc} = 1 if item i is classified in category 1 in group j, o if item i is classified in category 2 in group j

$Y_{ic} = \sum X_{ijc}$, the number of times a particular item is given a particular rating across groups of judges, (0 Y_{ic} J)

$P_{jc} = \frac{1}{I}\sum X_{ijc}$, the observed proportion of classifications into the c^{th} category by the j^{th} group

$Pc = \sum_{j} P_{jc/J}$, the overall proportion of classifications into the c^{th} category

Examples of how to compute Kappa using this formula with data from the current study are presented below.

The first example illustrated how the Kappa value of 0.75 presented in Table 2, page 73, for indicator I_3 for agreement among ethnic subgroup for word problems was determined. As a first step, the frequency of word problems judged biased by ethnic subgroups (black, hispanic, other) for Indicator I_3 must be examined. These data are presented for this purpose.

	B	H	O		B	H	O		B	H	O
Item 96	3	3	2	Item 106	1	1	0	Item 116	0	0	0
97	0	0	0	107	0	0	0	117	0	0	0
98	0	1	0	108	0	0	0	118	0	0	0
99	0	0	0	109	0	0	0	119	0	0	0
100	0	0	0	110	1	1	0	120	0	0	0
101	0	1	0	111	1	0	0	121	0	0	0
102	0	0	0	112	0	1	1	122	0	0	1
103	1	1	1	113	2	0	0	123	0	0	0
104	0	0	0	114	0	0	0	124	0	0	1
105	0	0	0	115	0	0	0	125	0	0	0
									$\frac{2}{30}$	$\frac{1}{30}$	$\frac{1}{30}$

For an item to be considered biased by a subgroup, according to the decision rule employed in this study, two or more judges within a subgroup had to identify the item as biased. Thus, examining the data above, two of the thirty word problems were considered biased by the black subgroup, one of the thirty word problems was considered biased by the white subgroup, and one of the thirty word problems was considered biased by the other subgroup.

By examining the data for each item it is possible to see how many of the ethnic subgroups judged an item to be biased. Again, for an item to be considered biased by a subgroup, two or more judges within a subgroup had to identify the item as biased. Item 96 was considered biased

by all three subgroups; items 97 through 112 were considered unbiased by all three subgroups; item 113 was considered biased by the black subgroup but not by the Hispanic or other subgroups; and items 114 through 125 were considered unbiased by all three subgroups. Thus, in inspecting the data, it can be seen that there was a great deal of agreement between the subgroups in their ratings of the items.

Kappa is computed using the formula developed by Davies and Fleiss (1982) presented above:

$$K = 1 - \frac{Y_{ic}(J-Y_{ic})}{I\left[J(J-1)\bar{P}_c(1-\bar{P}_c) + (p_{jc}-\bar{P}_c)^2\right]}$$

Note that the expression:

$$Y_{ic}(J-Y_{ic}) = 0 \text{ for } Y_{ic} = 0 \text{ or } 3$$
$$= 2 \text{ for } Y_{ic} = 1 \text{ or } 2$$

Therefore, using the above data, the equation becomes:

$$K = 1 - \frac{1.2}{30\left[3 \cdot 2 \cdot \frac{4}{90}\left(1 - \frac{4}{90}\right) + \left(\frac{2}{30} - \frac{4}{90}\right)^2 + 2\left(\frac{1}{30} - \frac{4}{90}\right)^2\right]}$$

$$= 1 - \frac{2}{30\left[3 \cdot 2 \cdot \frac{4}{90}\left(1 - \frac{4}{90}\right) + \left(\frac{6}{90} - \frac{4}{90}\right)^2 + 2\left(\frac{3}{90} - \frac{4}{90}\right)^2\right]}$$

$$= 1 - \frac{2}{30\left[\frac{6}{90}\ \frac{4}{90}\ \frac{86}{90} + \frac{4}{90} + \frac{2}{90}\right]}$$

$$= 1 - \frac{2}{\frac{30}{90}^2 \cdot 6 \cdot 4 \cdot 86 + 6}$$

$$= 1 - \frac{2}{8.94} -$$

$$= 1 - .25$$

$$K = .75$$

The Kappa statistic indicating significant agreement among ethnic subgroups for indicator I_3 for word problems is .75 ($x = .01$) as presented in Table 2.

A second example illustrates how the Kappa value of .07 also presented in Table 2 for indicator I_3 for agreement among ethnic subgroups for computational items was determined. Again, as a first step, the frequency of computational items judged biased by ethnic subgroups (black, Hispanic, other) for I_3 must be examined. These data are presented for this purpose.

	B	H	O		B	H	O		B	H	O
Item 1	1	0	0	Item 9	1	0	1	Item 17	0	1	0
2	1	0	1	10	1	0	0	18	0	1	0
3	0	0	1	11	1	0	1	19	0	1	0
4	0	0	1	12	0	1	0	20	0	1	1
5	0	1	0	13	0	1	0	21	1	1	1
6	0	1	0	14	0	1	0	22	0	1	0
7	0	0	0	15	0	1	0	23	2	1	2
8	0	0	0	16	0	1	0	24	0	0	0

Item				Item				Item			
25	0	1	0	51	0	2	0	77	0	0	0
26	2	1	2	52	0	2	0	78	0	0	0
27	0	2	0	53	1	1	0	79	0	0	0
28	1	2	0	54	0	0	0	80	0	0	1
29	1	2	1	55	0	1	0	81	0	0	0
30	1	0	2	56	0	0	0	82	0	0	1
31	1	1	2	57	0	0	0	83	0	1	1
32	0	0	2	58	0	1	0	84	0	2	1
33	0	0	2	59	0	1	0	85	0	2	1
34	2	1	3	60	0	1	0	86	0	1	1
35	0	0	0	61	0	1	0	87	0	1	1
36	0	0	3	62	0	2	0	88	0	0	0
37	1	0	1	63	1	2	0	89	0	1	1
38	0	0	0	64	1	2	0	90	0	0	0
39	0	0	1	65	1	0	0	91	0	0	2
40	0	0	1	66	1	0	2	92	1	0	0
41	0	0	0	67	0	0	1	93	0	0	2
42	0	0	0	68	0	0	2	94	0	2	3
43	0	2	0	69	2	1	3	95	1	2	3
44	0	0	0	70	1	0	3				
45	0	0	0	71	1	0	3		$\frac{5}{95}$	$\frac{16}{95}$	$\frac{17}{95}$
46	1	1	0	72	1	0	1				
47	1	1	0	73	0	0	0				
48	2	2	0	74	0	0	0				
49	1	2	0	75	0	0	0				
50	1	2	0	76	0	0	0				

In examining the data above, 5 of the 95 computational items were considered biased by the black subgroup; 16 of the items were considered biased by the Hispanic subgroup; and 17 of the items were considered biased by the other subgroup. Unlike the first example, an inspection of the data for individual items indicates a number of cases for which there was little agreement among the subgroups. These are the case where there is a two or greater for one subgroup, but not for the others.

Kappa is again computed using the formula described above:

$$K = 1 - \frac{Y_{ic}(J-Y_{ic})}{I\left[J(J-1)P_c(1-P_c) + (P_{jc}-P_c)^2\right]}$$

$$= 1 - \frac{31 \cdot 2}{95\left[(3 \cdot 2 \cdot \frac{38}{285})(1-\frac{38}{285}) + (\frac{5}{95}-\frac{38}{285})^2 + \frac{16}{(95}-\frac{38}{285})^2 + (\frac{17}{95}-\frac{38}{285})^2\right]}$$

$$= 1 - \frac{62}{95\left[(3 \cdot 2 \cdot \frac{38}{285})(1-\frac{38}{285}) + (\frac{15}{285}-\frac{38}{285})^2 + (\frac{48}{285}-\frac{38}{285})^2 + (\frac{51}{285}-\frac{38}{285})^2\right]}$$

$$= 1 - \frac{62}{95\left[(\frac{6\ 28\ 247}{285}) + (\frac{23}{285})^2 + (\frac{10}{285})^2 + (\frac{13}{285})^2\right]}$$

$$= 1 - \frac{62}{95\left[(\frac{6\ 28\ 247}{285}) + \frac{529}{285}^2 + \frac{100}{285}^2 + (\frac{169}{285})^2\right]}$$

$$= 1 - \frac{62}{\frac{95}{285}^2 = (6 \cdot 38 \cdot 247 + 798)}$$

$$= 1 - \frac{31}{\frac{1}{3.285}\ (28557)}$$

$$= 1 - \frac{1}{33.4}$$

$$= 1 - .928$$

$$K = .072$$

The Kappa statistic indicating no significant agreement among ethnic subgroups for Indicator I_3 for computational items is .07 as presented in Table 2.

When using Kappa to test agreement, the sampling variation should be taken into account by dividing Kappa by its estimated standard error and comparing the resulting standardized statistic to Z_{α}.

The test was rescored, eliminating those items that were determined to be biased for any subgroup based on statistical bias identification procedures. The effects of rescoring the test on the average subgroup scores were examined.

Items identified as biased by either or both method(s) were examined and compared. An effort was made to pinpoint possible sources of bias and to determine patterns that may be contributing to bias in terms of content, use of format, or use of language. The written explanations provided by judges as to why they rated particular items biased were used in this process.

Correlations of the existing test and of the rescored test, after eliminating biased items, with the mathematics subtest of the Stanford Test of Academic Skills (1971) were determined. These correlations were computed to examine bias in the existing test and the rescored test with respect to an external criterion (Ironson, 1980).

Results

The responses of twenty-four judges to three questions about each computational item and eight questions about each word problem on Test 2 provided data about perceived bias. Eight of the reviewers were black, eight of the reviewers were white, and eight of the reviewers were Hispanic. Within each group, half of the reviewers were male and half were female. These data were used to compute several indicators of bias as described above: I_3, I_8, and Q_1.

The administration of Test 2 to 1,064 high school students of both sexes and diverse ethnic backgrounds provided data on 121 items that were analyzed using the LOGIST program. Using the method of analysis described, an odd/even split of subjects resulted in the identification of ten items with significant z's at the .05 level. This is slightly larger than the number of significant z's expected by chance and reflects the fact that the test is not unidimensional. It provides an empirical baseline for the identification of biased items within the groups of interest. If the number of items identified as biased is greater than ten, it is likely that items are differentially more difficult for one subgroup.

The Kappa statistic was used to assess agreement between judgmental ratings among subgroups of reviewers, statistical ratings among subgroups of students, and agreement between item bias ratings by corresponding subgroup characteristics of judges and students. The results of these analyses in providing answers to the critical questions being investigated are discussed below.

1) Will the test item identified as biased by reviewers from different ethnic or gender subgroups be related across subgroups?

In investigating the agreement among reviewers who were asked to answer questions about each item to determine if bias existed in the item, several indicators of bias were created based on the reviewers responses to the questions. The indicators calculated were: I_3, Q_1, and Q_8. The procedure for determining these indices is described in the methodology section.

In Table 2, between subgroup agreement is presented using the Kappa statistic. For the ethnic and gender subgroups, agreement is presented separately for computational problems and word problems. For computational items, agreement is presented for indicators I_3 and Q_1. For the ethnic subgroups, the Kappa statistic for computational items is 0.07 for I_3 and 0.18 for Q_1, indicating that there is no significant agreement between ethnic subgroups on Q_1 or on I_3. For the gender subgroups, the Kappa statistic for

Table 2

Agreement Among Subgroups for Indicators I_3, Q_1, and Q_8 for Computational Problems and Word Problems for Ethnic and Gender Subgroups

Subgroups	Indicator	Kappa
Computational Problems		
Ethnic	I_3	0.07
	Q_1	0.18
Gender	I_3	0.23*
	Q_1	0.21**
Word Problems		
Ethnic	I_3	0.75**
	Q_1	0.00
	I_8	0.29**
Gender	I_3	0.25
	Q_1	0.00
	I_8	0.13

* $\alpha = .05$
** $\alpha = .01$

computational items is 0.23 for I_3 and 0.21 for Q_1, indicating significant agreement between gender subgroups on I_3 (at $\alpha = .05$) and for Q_1, (at $\alpha = .01$.)

For word problem items, agreement is presented for indicators I_3, Q_1, and I_8. (It should be noted that I_8 is an index pertaining only to word problems.) For the ethnic subgroups, the Kappa statistic for word problems is 0.75 for I_3, 0.00 for Q_1, and 0.29 for I_8. Thus, between ethnic subgroups, there is significant agreement among judges for I_3 (at $\alpha = .01$) and I_8 (at $\alpha = .05$), but not for Q_1. For the gender subgroups, the Kappa statistic for word problems is 0.03 for I_3, 0.00 for Q_1, and 0.13 for I_8. Thus, between gender subgroups, there were no significant agreement among judges on I_3, Q_1, or I_8.

In Table 3, the number of items identified as biased by judges ratings, using the procedures described, for ethnic status subgroups (black reviewers versus Hispanic reviewers versus other reviewers) and gender subgroups (male reviewers versus female reviewers) are presented. Data are presented for indicators I_3, Q_1, and I_8.

Based on the results of the above analyses, a decision was made not to use I_8 in further analyses. I_8 does not appear to be providing more information than I_3. Although I_8 contains information on more questions about each item than I_3, the value of I_8 was smaller than that of I_3 for both ethnic and gender subgroups. It should be noted that

Table 3

Number of Items Identified as Biased by Judges' Ratings for Ethnic Subgroup and Gender Subgroups According to Indicators I_3, Q_1, and I_8.

Ethnic			Gender		
I_3	Q_1	I_8	I_3	Q_1	I_8
33	9	21	33	21	28

I_8 could be calculated only for the word problems. I_8 was based on all eight questions asked about each word problem: three of these questions were questions that were asked about both computational problems and word problems; five of these questions were specific to word problems. Because of this, I_8 had the further disadvantage of not being usable across all items. Thus, the decision was made to drop I_8 from subsequent analyses.

The data presented in Table 4 are the agreement between judges' overall bias ratings using ethnic versus gender subgroups on indicators I_3 and Q_1. The overall bias ratings using ethnic subgroups represent bias recognized by at least two judges for any ethnic group and the overall bias ratings using gender subgroups represent bias recognized by at least two judges of either sex. The Kappa value for I_3 is 0.63 and the Kappa value for Q_1 is 0.48; both are significant at $\alpha = .001$. Thus there is agreement between judges from ethnic and gender subgroups about which items are biased. This agreement is probably largely artifactual and due to the relatively small number of judges, each of whom was in two different subgroups depending upon the analysis being conducted.

2. Will the test items identified as biased by LOGIST for ethnic subgroups of students be related to those identified as biased for gender subgroups of students?

Table 4

Agreement Between Judges' Overall Bias Ratings Using Ethnic Versus Gender Subgroups on Indicators I_3 and Q_1

Indicator	Kappa
I_3	0.63*
Q_1	0.48*

* $\alpha = .001$

The Kappa statistic indicating agreement between LOGIST bias ratings for ethnic and gender subgroups is not significant (K = 0.02). This indicates that the items on the test that are differentially more difficult between gender subgroups (for males versus females) are not the same as those that are differentially more difficult between ethnic subgroups (for blacks versus non-blacks).

Using LOGIST, a test of significance was performed on the b_g's for the black and non-black groups. This test resulted in thirty-one significant items, more than three times the number of significant items found on the odd/even split of students. Eleven of these items were identified as biased against blacks; twenty as biased against non-blacks. The test of significance was also performed on the b_g's for the male and female groups. Twenty-one significant items were identified, more than twice as many as in the odd/even split of students. Nine of these items were identified as biased against males; twelve as biased against females. These data are presented in Table 5.

The items that were significant in the black and non-black groups were not, for the most part, the same items that were found significant in the male and female groups. In Table 6, the items that were identified as biased by LOGIST for ethnic subgroups (black versus non-black students) and gender subgroups (male versus female students) are presented.

Table 5

Number of Items Identified as Biased

by LOGIST for Ethnic Subgroups and Gender Subgroups

Ethnic Subgroups		Gender Subgroups	
Blacks	Non-blacks	Males	Females
11	20	9	12

As can be seen from an examination of the data presented in Table 6, the distribution of items identified as biased against males or females does not show a pattern. However, this is not the case for items identified as biased against blacks or non-blacks. The items that were identified as biased against non-blacks were all within the first 65 items on the test; the items that were identified as biased against blacks were all items between items 66 and 125 on the test.

The first 95 items on the test are computational; the remaining 30 items are word problems. Items 1 through 31 test knowledge of whole numbers and items 32 through 65 test knowledge of fractions. Items 66 through 95 assess knowledge of decimals and percents. Items 95 through 125 test the use of these operations in the context of word problems. Items testing knowledge of whole numbers and fractions that were identified as biased by LOGIST were biased without exception against non-blacks; whereas items testing knowledge of decimals and percentages and word problems that were identified as biased were biased without exception against blacks.

3. Will statistical and judgmental methods for indentifying biased items on a test agree?

In Table 7, the agreements between statistical and judgmental rating (I_3 and Q_1) procedures in the identification of biased items for ethnic and gender subgroups are

Table 6

Items Identified as Biased by LOGIST for Ethnic Subgroups and for Gender Subgroups and the Subgroup Against Which the Item is Biased

	Ethnic Subgroups	Subgroups	Gender Subgroups	Subgroups
Item Numbers*	10	non-blacks	12	females
	11	non-blacks	22	males
	17	non-blacks	25	males
	18	non-blacks	31	females
	23	non-blacks	36	males
	35	non-blacks	44	males
	38	non-blacks	45	males
	44	non-blacks	68	females
	45	non-blacks	69	females
	46	non-blacks	70	females
	48	non-blacks	72	females
	51	non-blacks	73	males
	53	non-blacks	74	males
	55	non-blacks	75	females
	56	non-blacks	76	females
	57	non-blacks	80	males

*Items are presented in Appendix I.

Table 6 (continued)

Ethnic Subgroups	Subgroups	Gender Subgroups	Subgroups
58	non-blacks	83	males
59	non-blacks	90	females
60	non-blacks	100	females
61	non-blacks	111	females
66	blacks	112	
72	blacks		
82	blacks		
83	blacks		
84	blacks		
98	blacks		
103	blacks		
105	blacks		
106	blacks		
111	blacks		
112	blacks		

Table 7

Agreement Between Statistical and Judgmental Rating (I_3 and Q_1) Procedures for Bias Identification for Ethnic Gender Subgroups

Subgroups	Indicator	Kappa
Ethnic	I_3	0.12
	Q_1	0.04
Gender	I_3	0.04
	Q_1	0.03

presented. As can be seen from an examination of these data, there is no significant agreement between item bias detection methods. Kappa is not significant for either I_3 and LOGIST (K = 0.12) or Q_1 and LOGIST (K = 0.04) for ethnic subgroups; nor is Kappa significant for I_3 and LOGIST (K = 0.04) or Q_1 and LOGIST (K = 0.03) for gender subgroups. It appears that the statistical and the judgmental procedures are identifying different kinds of bias.

4. Will a test that is rescored eliminating items identified as biased yield significantly higher student outcome results for each subgroup and for the total group?

In order to rescore the test, only items identified as biased by LOGIST were used in determining which items to eliminate. Because there was such poor agreement between the items identified as biased by LOGIST and those identified as biased by the judgmental procedure (even given the very liberal decision rules employed), it was deemed most appropriate to eliminate only items identified by LOGIST in that these items were accounting for actual differences in performance between subgroups. Reviewers were responding to characteristics of items other than those which were responsible for differential performance as determined by LOGIST. The reviewers' responses and what they appear to have tapped are treated more fully in the Discussion section.

As presented in Table 6, thirty-one items were identified as biased by LOGIST for the ethnic subgroups (blacks and non-blacks) and twenty-one items were identified as biased by LOGIST for the gender subgroups (males and females). There is not significant agreement in items that were identified as biased using ethnic versus gender subgroups. Among the fifty-two items that were identified as biased for one characteristic (ethnicity or gender), only six items were identified as biased in both analyses. Thus, forty-six items were identified as biased for either ethnicity or gender. In rescoring Test 2, these forty-six items were eliminated. The item numbers are presented in Appendix II.

Test 2 was rescored using the 69 remaining items for each subgroup and for the total group. The proportion of items correct was determined for blacks, non-blacks, males, females, and for the total group.

In order to determine whether significant differences in average test results would occur for each subgroup and for the total group when biased items were eliminated, several correlated t-tests were performed comparing the proportion of items correct on Test 2 with the proportion of items correct on a rescored version of Test 2. T-values were determined for blacks, non-blacks, males, females, and for the total group. These results are presented in Table 8. In each case, significant differences occurred at the $\alpha = .001$

Table 8

Differences Between The Proportion of Items Correct on Test 2 and on Rescored Test 2 (Eliminating Biased Items) for Each Subgroup and for the Total Group

	Proportion of Items Correct on . . .		Difference: Proportion of Items Correct on Rescored Test 2 - Proportion of Items correct on Test 2	
	Rescored Test 2	Test 2		t
Blacks	0.73	0.70	0.03	13.19*
Non-blacks	0.75	0.72	0.03	13.69*
Males	0.74	0.71	0.02	12.06*
Females	0.74	0.71	0.03	15.20*
Total Group	0.74	0.71	0.03	19.44*

* $\alpha = .0001$

level. The t-values are 13.19 for blacks, 13.69 for non-blacks, 12.06 for males, 15.20 for females, and 19.44 for the total group. As can be seen from an examination of the data, the differences for each subgroup and for the total group are of the same order of magnitude. The effect of rescoring the test is to increase the average score by about two and one-half percent for each subgroup and for the total group. However, when the test is rescored, the rank order of subgroups and the total group does not change.

5. Will student performance on a criterion-referenced test in mathematics and on a rescored version of the test, eliminating items identified as biased, be correlated with student performance on the mathematics subtest of the Stanford Test of Academic Skills, Level 1 (1971) and will these two versions of the test be correlated with one another?

The correlations of student performance on Test 2 and a rescored version of Test 2, eliminating items identified as biased, with student performance on the mathematics subtest of the Stanford Test of Academic Skills, Level 1 (1971) are presented in Table 9. The correlation of student performance on Test 2 with student performance on a rescored version of Test 2 is also included in this table. The correlation of Test 2 with the Stanford Test of Academic

Table 9

The Correlations Between Student Performance on Test 2 and Rescored Test 2 With Student Performance on the Stanford Test of Academic Skills and the Correlation Between Test 2 and Rescored Test 2

Test 2 and the Stanford Test of Academic Skills	Rescored Test 2 and the Stanford Test of Academic Skills	Test 2 and Rescored Test 2
.32*	.32*	.96*

* α = .001

Skills is .32 and the correlation of the rescored version of Test 2 with the Stanford Test of Academic Skills is .32.

These correlations are virtually identical and they are both significant at the $\alpha = .001$ level. The correlation between student performance on Test 2 and the rescored version of Test 2 is extremely high: $r = .96$ ($\alpha = .001$).

Discussion

Two procedures for detecting biased items in tests were explored in this investigation. A statistical procedure and a judgmental procedure were employed in identifying items that might be labeled biased on a criterion-referenced mathematics test. The statistical procedure used was the three-parameter latent trait method (LOGIST) to identify items which were differentially more difficult for members of particular subgroups. This is considered to be one definition of bias. The judgmental procedure made use of twenty-four reviewers, of three different ethnic groups and both gender groups, who rated each item on the test for various potential sources of bias. The judges used a structured review form and answered questions about item difficulty, content, form, and language. The material to which judges responded primarily involved offensive activities, stereotyped descriptions, and a lack of representativeness of different ethnic and gender subgroups. This is considered to be another definition of bias. Agreement between the statistical procedure and the judgmental procedure in identifying biased items was assessed.

In order to investigate the use of statistical methods and judgmental methods in detecting biased items, and the

relationship between these two approaches, five critical questions were investigated:

1. Will the test items identified as biased by reviewers from different ethnic or gender subgroups be related across subgroups?

2. Will the test items identified as biased by LOGIST for ethnic subgroups of students be related to those identified as biased by gender subgroups of students?

3. Will statistical and judgmental methods for identifying biased items on a test agree?

4. Will a test that is rescored eliminating items identified as biased yield significantly higher student outcome results for each subgroup and for the total group?

5. Will student performance on a criterion-referenced test in mathematics and on a rescored version of the test, eliminating items identified as biased, be correlated with student performance on the mathematics subtest of the Stanford Test of Academic Skills, Level 1, (1971) and will these two versions of the test be correlated with one another?

In this section, the conclusions that may be drawn from the data collected in answering each of the critical questions will be discussed. This will be followed by a consideration of general issues raised in the study and suggested areas for further research.

1. Will the test items identified as biased by reviewers from different ethnic or gender subgroups be related across subgroups?

The reviewers identified very few items on the test as biased or difficult. For this reason, the liberal decision rule described in the methodology section to compute indicators I_3, I_8, and Q_1 was adopted. This made it possible to use most of the information from the rating data provided by the reviewers. The large majority of the reviewers' written comments, collected in addition to their ratings, pertained to the word problems rather than to the computational problems.

As the data presented in Table 2 indicates, there was significant agreement between the judges on some of the indicators and not on others. An inspection of these results does not indicate a pattern that is uniform or consistent. For ethnic subgroups, there was significant agreement among the judges on two of the indicators (I_3 and I_8) for word problems, but no significant agreement on the indicators for computational problems. The agreement among gender subgroups was reversed. There was significant agreement among the judges for computational problems on the two relevant indicators (I_3 and Q_1), but no significant agreement among the judges on any of the indicators for word problems.

Because of the relatively few items that received biased ratings from reviewers from any of the subgroups, it would be unwarranted to attempt to interpret a pattern among the judges' responses. Furthermore, all of the reviewers had two salient subgroup characteristics, each of which was used to designate a reviewer as a member of one subgroup or another according to the analysis being conducted. That is, for one analysis the judges were divided into subgroups by ethnicity (black, Hispanic, or other) and for another analysis the same judges were divided into subgroups by gender (male and female). In this way, the ratings from each reviewer were used twice; once for the ethnic subgroups analysis and once for the gender subgroups analysis. The highly significant agreement between reviewers, presented in Table 3, is largely artifactual and due to this division of judges. In order to meaningfully explore patterns among the ratings of judges, it would be necessary to have data on a test on which there was a fairly large number of items perceived as biased. However, the results of the present investigation suggest that items on recently developed tests may be subjected to ongoing screening throughout the test construction process. As a result, few items are perceived as biased on the final version of the test.

The judges' comments about the problems that they rated as biased, particularly the word problems, are of interest. When responding to these problems, reviewers cited activi-

ties and descriptions that they believed might be offensive or unfamiliar to particular subgroups, language that they believed might be offensive or unfamiliar to particular subgroups, and, to a lesser extent, structure or format of the problem that they believed was generally difficult. Several examples of reviewers' comments to particular test items will serve to illustrate their concerns:

Item 96 read:

> "There were 12,000 spark plugs produced yesterday at a factory. They were packed in boxes of 6 spark plugs each. How many boxes were packed? Which expression below cannot be used to solve the problem correctly?"

A Hispanic male reviewer commented:

> "The spark plug expression may be unknown for certain students. If the problem is stated, using candies for example, the students will have a better chance to direct their efforts to the solution."

A white male reviewer stated:

> "A spark plug is possibly a gender related (i.e., recognized by men but not necessarily by women) object.

Thereby possibly confusing and discriminating against women."

A white female reviewer remarked about the difficulty of this problem (unlike most of the comments about word problems):

"Students must find the answer to the problem by selecting the answer choice that does not answer the question. The sentence structure 'cannot be used to solve the problem' is difficult."

Item 98 read:

"Jose can paint an apartment in 12 hours and Maria can do the same job in 9 hours. How much less time does it take Maria to do the job?"

A black woman reviewer commented:

"It would be most appropriate and realistic to use two people of the same sex. Using different sexes may imply a competition between the sexes."

Another black woman reviewer stated:

"Both sexes are not equally portrayed. The problem should

> mention either two boys' or two girls' names, not one of each."

A Hispanic male reviewer stated:

> "This may be an offensive comparison for males, especially Hispanic males. Why not Jose and Pedro or Rosa and Maria?"

A white male reviewer believed the problem was ethnically and gender offensive, commenting:

> "Both individuals have Hispanic names and are involved in menial labor. More generic names could be used. The woman is better or quicker than the man. This is potentially insulting."

A Hispanic male reviewer who commented negatively about the problem also stated:

> "There are fifteen word problems on the test including persons and only this one has (latin) Hispanic names."

Judges, regardless of their subgroup characteristic (i.e., black, Hispanic, white, female or male), seemed to make the same kinds of comments when they identified a problem as offensive. Members of a particular subgroup did not appear to be more sensitive to concerns of their own

subgroup than to others. It should be noted again that the number of bias ratings per judge was relatively small and unwarranted generalizations should not be made from the finding.

2. Will the items identified as biased by LOGIST for ethnic subgroups of students be related to those identified as biased by gender subgroups of students?

There was no significant agreement between items identified as biased by LOGIST for ethnic and gender subgroups of students. As reported above, there were six common items of the 52 items that were identified as biased for either ethnic or gender subgroups. These findings reflect the fact that LOGIST is detecting factors other than pure mathematics ability which are effecting student performance on these items. For males and females, these factors appear to have been randomly distributed throughout the test. This was not the case, however, for black and non-blacks.

The items identified as biased against non-blacks were computational items which tapped knowledge of whole numbers and fractions. The items identified as biased against blacks were computational items which tapped knowledge of decimals and percentages and word problems. A number of explanations might be proposed for these findings.

The operations on the test, and the corresponding computational items, were organized by progressive difficulty

according to the mathematics experts who were part of the test development process. Items involving whole numbers were considered to be easier than those involving fractions. Items involving fractions were considered to be easier than those involving decimals and percents. Word problems tapped one or more operations and may have been more difficult for this reason, as well as because of the use of language. Some of the computational items also included language -- usually short declarative phrases or questions. Most of the computational items which included a phrase or a sentence were among the items which tapped knowledge of decimals and percents.

Thus, the latter part of the test, word problems and many items involving decimals and percents, was measuring factors other than mathematics ability. It was probably tapping language skills and reading achievement, as well as other areas. These areas may well be areas in which black students have had less "opportunity to learn" (Cooley and Leinhardt, 1978). In addition, the language which black students have learned may not be standard English, so the words and phrases in the test may be less familiar to them than to other students.

It should be noted again that while there was not significant agreement between the statistical procedure and the judgmental procedure, most of the judges' comments were about the word problems. One reviewer, a black woman,

commented on the language used in many of the word problems and her concerns reflect those expressed by many of the judges.

Item 99 read:

> "Fran sold 50 records and 150 tapes at a flea market. How many items did she sell?"

The reviewer commented:

> "Although the phrase 'at a flea market' does not have to be understood to solve the problem, a few poor readers may ponder over it thinking that it might be important. Why include this phrase at all?"

Four other reviewers also commented on the use of this phrase.

Item 110 read:

> "In a room requiring 16 square yards of carpeting, 5-3/4 square yards have already been installed. How many yards of carpet remain to be installed?"

The same reviewer stated:

> "Requiring = needing, installed = put in. These changes make the item easier to read for those not familiar with the vocabulary or with having carpeting installed."

Item 111 read:

> "Calvin helps in a grocery store 4-1/2 hours a day for six days a week. How many hours a week does he work in the grocery store?"

The same reviewer commented:

> "Once again, many students do not distinguish between important words and those which make no difference to the problem. Why not take out the word 'grocery' to avoid such a situation?"

Three other reviewers commented about the language or description in this item. It was also one of the items identified by LOGIST as being biased against black students.

The use of difficult or unfamiliar language in the word problems was noted by many of the reviewers. It may explain why the word problems that were identified as biased by LOGIST were biased against black students, although the judges did not, for the most part, cite the same items as LOGIST. Similarly, the computational items that were identified as biased against blacks were, in most cases, items that contained written phrases or questions.

None of the reviewers' comments provides an explanation as to why the computational items tapping whole numbers and fractions that were identified as biased by LOGIST were biased against non-black students. It may be that black students, many of whom have received years of remediation in basic skills, are more familiar with basic operations than

non-black students. Non-black students may have received less remediation in basic skills and may make more computational errors than black students and, thus, do relatively less well on these items. More research is needed to investigate these differences in performance between black and non-black students.

3. Will statistical and judgmental methods for identifying biased items on a test agree?

No significant agreement was found between items identified as biased by LOGIST and those identified by the reviewers. LOGIST identified different biased items for ethnic and gender subgroups, suggesting a fairly random distribution of differentially difficult items. Reviewers rated relatively few items as biased. The comments of the reviewers suggest that for the most part they were responding to something quite different than the differential difficulty of the items that was identified by LOGIST.

Eleven of the reviewers commented about Item 96, which is presented above as an example. Many of them believed that spark plugs was a term that might be unfamiliar to certain subgroups, particularly women. Item 96 was not, however, identified by LOGIST as a differentially more difficult problem for gender subgroups (i.e., females versus males) or for ethnic subgroups (blacks versus non-blacks). While the judges were responding to a term that might be confusing, it did not adversely effect the performance of

any subgroup of students. It should be noted that knowledge of the term was not necessary to solve the problem. It may be that some students did not know the term, but were test-wise enough not to be confused by it and to realize that they could solve the problem anyway.

Item 98, also provided as an example above, is an anomaly in several ways. It is one of the problems on the test where there was some agreement between reviewers and LOGIST. It is the only problem in which a woman is involved in any activity that might traditionally be considered a man's job, and she does the job more quickly than the man to whom she is compared in the problem. It is also the only problem on the test in which Hispanic names are used. Ten reviewers commented on this problem, citing: the potential offensiveness of a "competition between the sexes," the potential offensiveness of a women doing a job more quickly than a man, the potential offensiveness of menial work being done by persons with Hispanic names, and possible difficulty with the problem because of the wording of the question. This problem was also identified as biased by LOGIST for ethnic subgroups, although not for gender subgroups.

4. Will a test that is rescored eliminating items identified as biased yield significantly higher student outcome results for each subgroup and for the total group?

When those items that were identified as biased by LOGIST were eliminated from the test, there were significant

increases in average scores for each subgroup and for the total group. These increases were uniform. The average increase for each subgroup and for the total group was about two and one-half percent. The deletion of biased items from the test did not make the test differentially less difficult for any subgroup and the relationship of the test to an external criterion (the mathematics subtest of the Stanford Test of Academic Skills, Level 1, (1971)) also remain unchanged. Thus, rescoring the test, with the biased items eliminated, probably made the test somewhat more unidimensional; that is, more dependent on the central concept being measured. However, it did not change the relative standing in terms of performance of any of the subgroups.

5. Will student performance on a criterion-referenced test in mathematics and on a rescored version of the test, eliminating items identified as biased, be correlated with student performance on the mathematics subtest of the Stanford Test of Academic Skills, Level 1 (1971) and will the two versions of the test be correlated with one another?

The correlations of the test and the rescored version of Test 2 with the Stanford Test of Academic Skills are virtually identical. Both correlations are fairly low, suggesting that the criterion-referenced test and the rescored version of the test are measuring considerably different skills than those measured by the Stanford Test

of Academic Skills. The very high correlation between Test 2 and the rescored version of Test 2, and the fact that the correlation with the external criterion remained nearly identical when such a large number of items was removed (about 40% of the original items), suggest that items that were removed were not very important in defining the underlying construct.

In summary, the current investigation was undertaken to explore methods of ensuring that a test is free of bias that discriminates against a particular subgroup of the population. Statistical procedures identify items that are differentially more difficult for one subgroup of the population than another. Psychometrically, an item that is significantly harder for one subgroup of the population than for other subgroups may be defined as biased. Statistical techniques, such as LOGIST, provide methods for identifying these items. In using these methods, data are also gathered about the balance of items on a test that are biased against different subgroups. This provides information about the content and construct validities of the test.

There were a number of problems encountered due to lack of fit of the data to the model when LOGIST procedures were employed. These problems are attributed to the multidimensionality of test, but raise questions about the use of latent trait methods. In the present study an effort was made to reduce the difficulties caused by the multi-

dimensionality of the test by analyzing discrete sections separately. This did not reduce the difficulties caused by the multidimensionality, but appeared to increase them. Decreasing the number of items by looking at discrete sections of the test made the remaining minor factors relatively more important. Most tests are not undimensional. The test that was revised in the present investigation was a criterion referenced mathematics test. It might be expected that the problem of multidimensionality would be even greater on a reading achievement test where more skills and processes are simultaneously assessed.

Latent trait methods need to be explored further with multidimensional tests with larger item pools to determine the fit between the model and data from real life test results. Procedures for handling multidimensionality need to be established for the three-parameter latent trait method to be more useful.

Judgmental procedures must be used to collect information about individual items, and overall test balance, that reviewers find offensive or stereotypical in their portrayal of a subgroup or subgroups of the population. These concerns constitute another definition of bias. In the present investigation the judgmental procedure did not yield many items that were identified as biased. In the items that were rated as biased, reviewers in this study were quite consistent in identifying items containing activities and

descriptions in which subgroups were portrayed in an offensive or stereotypical manner. They were also consistent in their comments about individual items and the lack of overall balance in the descriptions of activities in which different ethnic and gender subgroups were involved.
The data collected in this study indicated that the overall performance of subgroups of students was not affected by the items which reviewers identified as offensive. Furthermore, the data indicated that reviewers did not identify the differentially difficult items that were identified by the LOGIST procedure.

Although reviewers did not identify the same items as biased as were identified by LOGIST, many of the reviewers commented about the language used in the word problems and in several of the computational items. Reviewers suggested that these items included vocabulary or activities that may be unfamiliar to students from particular subgroups. The judges indicated that the vocabulary or activities might confuse these students. As was discussed above, differences in familiarity with language may provide an explanation for the differences in performance between black and non-black students. More research is needed in this area.

In that statistical procedures exist to identify differentially difficult items, it is not necessary that judges' ratings pinpoint these items. The role of judges should be to respond to aspects of test items which cannot

be identified statistically. Statistical methods and judgmental methods for identifying bias in items, and in tests, can be used to provide data about different fairness issues. Judgmental procedures can be used formally and informally at many stages of test development to ensure that a test does not contain a disproportionate number of items that portray a given subgroup in a manner that is stereotypical or offensive and statistical procedures can be used to ensure that a test does not contain a disproportionate number of items that are differentially more difficult for a given subgroup.

As was discussed above, the present investigation suggests that formal procedures for reviewing tests for biased items may be less essential with recently developed tests than with older tests. With newer tests, biased items may be eliminated during the test construction process. Results from the present study suggest the possibility that test developers have become increasingly sensitive to the concerns of different subgroups of the population about the fairness of tests and test items. This increased sensitivity has resulted, it appears, in an ongoing effort to ensure that items that might be considered offensive are not included on the final version of a test. More research is needed to investigate informal review of test items, the frequency with which informal review occurs, and the success of informal review in eliminating biased items in the construction of different tests.

The present investigation was conducted with students in remedial classes. Research is needed to determine whether restriction of ability range affected the results and whether the results are generalizable beyond this group.

APPENDIX I

THE NEW YORK CITY CRITERION-REFERENCED MATHEMATICS TEST

Computation Problems

Part I: Operations with Whole Numbers

1. Add

86 + 8 + 1985 =

A) 1979
B) 2049
C) 2072
D) 2079

2. Add

75 + 3804 + 821 =

A) 3800
B) 4530
C) 4700
D) 5390

3. Add

814
39
923
6837

A) 7613
B) 8513
C) 8603
D) 8613

4. Add

753
824
1072
95

A) 1534
B) 2644
C) 2735
D) 2744

5. Subtract

49
- 3

A) 36
B) 45
C) 46
D) 47

6. Subtract

89
- 64

A) 15
B) 24
C) 25
D) 35

7. Subtract

84
- 47

A) 37
B) 38
C) 43
D) 47

8. Subtract

381
- 92

A) 228
B) 289
C) 282
D) 299

9. Subtract

7006
- 279

A) 6727
B) 6737
C) 6837
D) 7279

10. Subtract

3075
- 296

A) 2779
B) 2789
C) 2821
D) 2879

11. Subtract

9000
- 7123

A) 1877
B) 1887
C) 1987
D) 2877

12. Multiply

41
X 2

A) 42
B) 43
C) 84
D) 82

13. Multiply

91
X 6

A) 156
B) 546
C) 654
D) 96

14. Multiply

795
X 4

A) 3160
B) 3432
C) 3180
D) 2880

15. Multiply

2146
X 7

A) 14,782
B) 14,982
C) 14,722
D) 15,022

16. Multiply $\begin{array}{r} 24 \\ \times\ 12 \\ \hline \end{array}$

A) 242
B) 72
C) 288
D) 212

17. Multiply $\begin{array}{r} 87 \\ \times\ 51 \\ \hline \end{array}$

A) 4437
B) 4427
C) 4137
D) 522

18. Multiply $\begin{array}{r} 517 \\ \times\ 23 \\ \hline \end{array}$

A) 11,771
B) 11,891
C) 11,871
D) 12,071

19. Multiply $\begin{array}{r} 324 \\ \times\ 20 \\ \hline \end{array}$

A) 64,800
B) 648
C) 344
D) 6,480

20. Multiply $\begin{array}{r} 241 \\ \times\ 201 \\ \hline \end{array}$

A) 2,241
B) 48,441
C) 5,061
D) 482,241

21. Multiply $\begin{array}{r} 300 \\ \times\ 112 \\ \hline \end{array}$

A) 33,600
B) 336
C) 3,360
D) 336,000

22. Divide $7\overline{)91}$

A) 29
B) 13
C) 23
D) 27

23. Divide $8\overline{)661}$

A) 79 R 28
B) 83
C) 82 R 5
D) 85 R 2

24. Divide $14\overline{)42}$

A) 8
B) 4
C) 3
D) 2

25. Divide $36\overline{)828}$

A) 32
B) 28
C) 23
D) 38

26. Divide $43\overline{)752}$

A) 15 R 7
B) 17 R 21
C) 17 R 11
D) 15 R 13

27. Divide $70\overline{)6300}$

A) 90
B) 99
C) 900
D) 9

28. Divide $12\overline{)24036}$

A) 23
B) 2003
C) 203
D) 230

29. Divide $21\overline{)2142}$

A) 120
B) 12
C) 10 R 2
D) 102

30. Which number represents forty thousand two hundred?

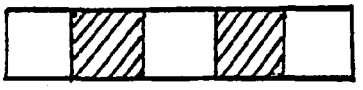
A) 4,020
B) 40,020
C) 40,200
D) 42,000

31. In the number 2,058 what is the value of the 5?

A) 5
B) 50
C) 500
D) 5 tenths

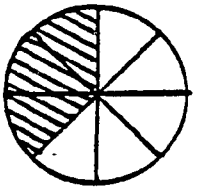
Part II: Fractions

32. The rectangle below is divided into equal parts. Which fraction describes the parts of the figure that are shaded?



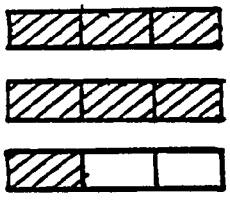
A) $\frac{2}{5}$
B) $\frac{3}{5}$
C) $\frac{2}{3}$
D) $\frac{5}{2}$

33. The circle below is divided into equal parts. Which fraction describes the parts of the figure that are shaded?



A) $\frac{8}{3}$
B) $\frac{3}{5}$
C) $\frac{5}{8}$
D) $\frac{3}{8}$

34. The rectanges below are divided into equal parts. Which number describes the parts of the figures that are shaded?



A) $3\frac{1}{3}$
B) $3\frac{2}{3}$
C) $2\frac{1}{3}$
D) $2\frac{1}{9}$

35. Which is larger, $\frac{2}{5}$ or $\frac{4}{5}$?

A) $\frac{2}{5}$
B) $\frac{4}{5}$

36. $\frac{2}{3} = \frac{\square}{15}$

What number does $\square$ represent?

A) 10
B) 14
C) 30
D) 5

37. What is $\frac{7}{3}$ written as a mixed number?

A) $2\frac{2}{3}$
B) $2\frac{1}{7}$
C) $3\frac{1}{3}$
D) $2\frac{1}{3}$

38. $5\frac{3}{4}$ is the same as

A) $\frac{53}{4}$
B) $\frac{23}{4}$
C) $\frac{15}{4}$
D) 23

39. What is $\frac{6}{9}$ reduced to lowest terms?

A) $\frac{1}{6}$
B) $\frac{3}{6}$
C) $\frac{2}{3}$
D) $\frac{9}{6}$

40. What is $\frac{2}{10}$ reduced to lowest terms?

A) $\frac{1}{10}$
B) $\frac{2}{5}$
C) $\frac{1}{2}$
D) $\frac{1}{5}$

41. Add $\frac{3}{6} + \frac{2}{6} =$

A) $\frac{5}{36}$
B) $\frac{1}{6}$
C) $\frac{5}{12}$
D) $\frac{5}{6}$

42. Add $\frac{1}{10} + \frac{2}{10} =$

A) $\frac{3}{100}$
B) $\frac{1}{10}$
C) $\frac{3}{20}$
D) $\frac{3}{10}$

43. Add $3\frac{3}{8} + 5\frac{2}{8}$

A) $8\frac{5}{8}$
B) $8\frac{5}{16}$
C) $8\frac{5}{64}$
D) $15\frac{5}{8}$

44. Add $\frac{1}{2} + \frac{1}{3} =$

A) $\frac{2}{5}$
B) $\frac{5}{12}$
C) $\frac{5}{6}$
D) $\frac{1}{6}$

45. Add $\frac{2}{3} + \frac{1}{4} =$

A) $\frac{3}{7}$
B) $\frac{3}{12}$
C) $\frac{1}{6}$
D) $\frac{11}{12}$

46. Add $11\frac{1}{4} + 7\frac{1}{8}$

A) $18\frac{3}{8}$
B) $18\frac{1}{16}$
C) $4\frac{1}{8}$
D) $18\frac{1}{32}$

47. Add $3\frac{5}{7} + 2\frac{3}{7}$

A) $5\frac{7}{8}$
B) $6\frac{1}{7}$
C) $5\frac{15}{49}$
D) $5\frac{1}{7}$

48. Add $5\frac{2}{3} + 3\frac{4}{5}$

A) $9\frac{6}{15}$
B) $8\frac{6}{8}$
C) $8\frac{7}{15}$
D) $9\frac{7}{15}$

49. Subtract $6 - 1\frac{1}{4}$

A) $4\frac{1}{4}$
B) $4\frac{3}{4}$
C) $5\frac{1}{4}$
D) $5\frac{3}{4}$

50. Subtract $5 - 2\frac{2}{3}$

A) $2\frac{1}{3}$
B) $2\frac{2}{3}$
C) $3\frac{1}{3}$
D) $3\frac{2}{3}$

51. Subtract $5\frac{1}{3} - 3\frac{2}{3}$

A) $1\frac{1}{3}$
B) $1\frac{2}{3}$
C) $2\frac{1}{3}$
D) $2\frac{2}{3}$

52. Subtract $6\frac{3}{5} - 5\frac{4}{5}$

A) $\frac{1}{5}$
B) $\frac{4}{5}$
C) $1\frac{1}{5}$
D) $1\frac{4}{5}$

53. Subtract $12\frac{1}{4} - 5\frac{1}{3}$

A) $6\frac{1}{12}$
B) $6\frac{11}{12}$
C) $7\frac{11}{12}$
D) $7\frac{1}{12}$

54. Multiply $\frac{3}{5} \times \frac{2}{7} =$

A) $\frac{5}{12}$
B) $\frac{6}{35}$
C) $\frac{5}{35}$
D) $\frac{2}{7}$

55. Multiply $\frac{3}{4} \times \frac{2}{5} =$

A) $\frac{3}{10}$
B) $\frac{5}{20}$
C) $\frac{3}{20}$
D) $\frac{8}{15}$

56. Multiply $1\frac{1}{6} \times 1\frac{5}{7} =$

A) $1\frac{5}{42}$
B) 2
C) $2\frac{5}{42}$
D) 3

57. Multiply $1\frac{1}{2} \times 2\frac{3}{4} =$

A) $2\frac{3}{8}$
B) $3\frac{3}{8}$
C) $4\frac{3}{8}$
D) $4\frac{1}{8}$

58. Multiply $5 \times 1\frac{1}{3} =$

A) $1\frac{1}{3}$
B) $2\frac{1}{3}$
C) $5\frac{1}{3}$
D) $6\frac{2}{3}$

59. Multiply $3 \times 1\frac{3}{4} =$

A) $1\frac{3}{4}$
B) $2\frac{1}{2}$
C) $3\frac{3}{4}$
D) $5\frac{1}{4}$

60. Divide $11 \div \frac{1}{3} =$

A) $\frac{3}{11}$
B) $3\frac{2}{3}$
C) $11\frac{1}{3}$
D) 33

61. Divide $7 \div \frac{3}{4} =$

A) $5\frac{1}{4}$
B) $7\frac{3}{4}$
C) $9\frac{1}{3}$
D) 21

62. Divide $\frac{1}{12} \div 2 =$

A) $\frac{1}{6}$
B) 24
C) 6
D) $\frac{1}{24}$

63. Divide $\frac{3}{5} \div 5 =$

A) 3
B) $\frac{3}{25}$
C) $\frac{1}{3}$
D) $8\frac{1}{3}$

64. Divide $7\frac{1}{2} \div 1\frac{1}{3} =$

A) $5\frac{5}{8}$
B) 10
C) $8\frac{2}{5}$
D) $4\frac{3}{8}$

65. Divide $1\frac{1}{5} \div 3\frac{2}{3} =$

A) $3\frac{2}{25}$
B) $4\frac{2}{25}$
C) $2\frac{1}{2}$
D) $\frac{6}{17}$

Part III: Decimals and Percents

66. Round 236 to the nearest ten.

A) 24

B) 200

C) 230

D) 240

67. Round 21,747 to the nearest thousand.

A) 21,700

B) 21,750

C) 21,800

D) 22,000

68. Round 958 to the nearest hundred.

A) 900

B) 950

C) 960

D) 1000

69. What number represents 27 thousandths?

A) 1027

B) .027

C) 27,000

D) .27

70. Round 5.80932 to the nearest hundredth.

A) 5.8

B) 5.80

C) 5.809

D) 5.81

71. Round 341.642 to the nearest hundredth.

A) 300

B) 341.6

C) 341.64

D) 341.65

72. Round 3.46 to the nearest tenth.

A) 3.0

B) 3.4

C) 3.45

D) 3.5

73. $\frac{2}{5}$ is the same as

A) .04

B) .25

C) .4

D) 2.5

74. $\frac{7}{4}$ is the same as

A) .175

B) .74

C) 1.75

D) 7.4

75. .03 is the same as

A) $\frac{3}{10}$

B) $\frac{3}{100}$

C) $\frac{10}{3}$

$\frac{100}{3}$

76. Add 4.72 + .003 + 12 =

A) 4.843

B) 4.87

C) 16.723

D) 16.75

77. Add

7.232 + .12 + 10.3 =

A) 8.382

B) 9.462

C) 17.652

D) 17.682

78. Subtract

2.04 - .72 =

A) 1.72

B) 1.32

C) 2.72

D) 2.76

79. Subtract

4.1 - .06 =

A) 3.5

B) 4.04

C) 4.14

D) 4.16

80. Multiply

.3 x .2 =

A) .006

B) .06

C) .6

D) 6

81. Multiply

.6 x .7 =

A) .0042
B) .042
C) 4.2
D) 42

82. Multiply

.04 x 2.2 =

A) .088
B) .88
C) 8.8
D) 88

83. Divide

$.05\overline{)3.65}$

A) .73
B) 7.3
C) 73
D) 730

84. Divide

$.6\overline{)1.86}$

A) .31
B) 3
C) 3.1
D) 31

85. Divide

$.16\overline{)3.2}$

A) .02
B) .2
C) 2
D) 20

86. Divide

$7\overline{).42}$

A) .6
B) .06
C) 6
D) 60

87. Divide $.5\overline{)1}$

A) .2
B) 2
C) .02
D) .5

88. 10% is the same as

A) 10
B) .10
C) 1.0
D) .01

89. .28 is the same as

A) 2.8%
B) 28%
C) .28%
D) .0028%

90. 7% is the same as

A) $\frac{7}{100}$
B) $\frac{1}{7}$
C) $\frac{7}{10}$
D) 7

91. 48% is the same as

A) 48
B) $\frac{10}{48}$
C) $\frac{12}{25}$
D) $\frac{12}{20}$

92. $\frac{4}{5}$ is the same as

A) 45%
B) 80%
C) 4.5%
D) 8%

93. What is 20% of 60?

A) .12
B) 1.2
C) 12
D) 1200

94. What % of 30 is 6?

A) 18%
B) 20%
C) 24%
D) 5%

95. 4% of a number is 20. Find the number.

A) 50
B) 80
C) 500
D) 5

THE NEW YORK CITY CRITERION-REFERENCED MATHEMATICS TEST

Word Problems

Part IV: Word Problems

Directions: In problems 96 to 99 choose the expression which would give the correct answer to the problem.

96. There were 12,000 spark plugs produced yesterday at a factory. They were packed in boxes of 6 spark plugs each. How many boxes were packed? Which expression below cannot be used to solve the problem correctly?

A) $12{,}000 \div 6$

B) $6\overline{)12{,}000}$

C) $6 \div 12{,}000$

D) $\frac{12{,}000}{6}$

---

97. If pens cost 19¢ each, then to find the cost of 6 pens we:

A) Multiply 6 by 19¢

B) Divide 6 by 19¢

C) Add 6 and 9¢

D) Subtract 6 from 19¢

---

98. Jose can paint an apartment in 12 hours and Maria can do the same job in 9 hours. How much less time does it take Maria to do the job?

A) $12 + 9$

B) $12 - 9$

C) $12 \div 9$

D) 12×9

---

99. Fran sold 50 records and 150 tapes at a flea market. How many items did she sell?

A) $150 - 50$

B) $150 \div 50$

C) $150 + 50$

D) 150×50

Directions: In problems 100 to 104 solve each and choose the correct answer.

100. 20 campers drank a total of 40 cups of milk this morning. Each camper drank the same amount. How many cups of milk did each camper drink?

A) 3

B) 60

C) 2

D) 20

---

101. Diane earned $35 every day for 14 days of work. How much money did she earn?

A) $350

B) $49

C) $490

D) $175

---

102. Alice had $115. She gave $32 to Sonia to buy a pair of shoes. How much did Alice have left to spend on a coat?

A) $115

B) $147

C) $83

D) $23

---

103. On Monday, the school store sold 18 chocolate bars and on Tuesday it sold 36 chocolate bars. How many chocolate bars were sold in those two days?

A) 54

B) 16

C) 2

D) 56

---

104. A high school basketball team scored 52, 48, 42 and 38 points in four home games. What was the average number of points scored?

A) 54

B) 90

C) 45

D) 180

Directions: In problems 105 to 108 choose the expression which would give the correct answer.

105. Pierre was $67\frac{3}{8}$ inches tall last year. Now he is $1\frac{1}{2}$ inches taller. To find out how tall he is now, we:

A) Add $1\frac{1}{2}$ to $67\frac{3}{8}$

B) Subtract $1\frac{1}{2}$ from $67\frac{3}{8}$

C) Multiply $67\frac{3}{8}$ by $1\frac{1}{2}$

D) Divide $67\frac{3}{8}$ by $1\frac{1}{2}$

106. John weighs $185\frac{1}{3}$ pounds. Phil weighs $162\frac{1}{2}$ pounds. To find out how much more John weighs than Phil weighs, we calculate:

A) $185\frac{1}{3} + 162\frac{1}{2}$

B) $185\frac{1}{3} - 162\frac{1}{2}$

C) $185\frac{1}{3} \times 162\frac{1}{2}$

D) $185\frac{1}{3} \div 162\frac{1}{2}$

107. Jerry had a bag of candy which weighed $\frac{3}{4}$ of a pound. He gave $\frac{1}{2}$ of it to Pat. To find out how much candy Jerry had left, we calculate:

A) $\frac{1}{2} + \frac{3}{4}$

B) $\frac{3}{4} - \frac{1}{2}$

C) $\frac{1}{2} \times \frac{3}{4}$

D) $\frac{3}{4} \div \frac{1}{2}$

108. Sue bought $42\frac{1}{2}$ yards of material to make costumes. Each costume requires $2\frac{1}{2}$ yards of material. To find out how many costumes she can make, we:

A) Add $42\frac{1}{2}$ to $2\frac{1}{2}$

B) Subtract $2\frac{1}{2}$ from $42\frac{1}{2}$

C) Multiply $2\frac{1}{2}$ by $42\frac{1}{2}$

D) Divide $42\frac{1}{2}$ by $2\frac{1}{2}$

Directions: In problems 109 to 112 solve each and choose the correct answer.

109. Harry works at a garage after school. On Monday he worked $2\frac{1}{2}$ hours, on Tuesday $3\frac{1}{4}$ hours. How long did he work those 2 days?

A) $5\frac{2}{6}$ hours

B) $5\frac{3}{4}$ hours

C) $\frac{3}{4}$ hour

D) $1\frac{1}{4}$ hours

110. In a room requiring 16 square yards of carpeting, $5\frac{3}{4}$ square yards have already been installed. How many square yards of carpet remain to be installed?

A) $21\frac{3}{4}$

B) $11\frac{1}{4}$

C) $10\frac{1}{4}$

D) $10\frac{3}{4}$

111. Calvin helps ina grocery store $4\frac{1}{2}$ hours a day for 6 days a week. How many hours a week does he work in the grocery store?

A) $1\frac{1}{2}$ C) 27

B) $10\frac{1}{2}$ D) 36

---

112. On a map, if $\frac{1}{8}$ inch represents 1 mile, how many miles will 6 inches represent?

A) $\frac{3}{4}$ C) $6\frac{1}{8}$

B) 48 D) 14

Directions: In problems 113 to 116 choose the expression which would give the correct answer.

---

113. A car gets 14.6 miles on a gallon of gas. To find out how far the car can travel on 7.3 gallons of gas. we calculate:

A) 14.6 + 7.3 C) 14.6 x 7.3

B) 14.6 - 7.3 D) 14.6 ÷ 7.3

---

114. If 6 chocolate bars cost \$1.62, to find out the cost of 1 chocolate bar, we calculate:

A) 1.62 + 6 C) 1.62 x 6

B) 1.62 - 6 D) 1.62 ÷ 6

---

115. John worked 6.5 hours on Saturday and 4.75 hours on Sunday. To find the total number of hours he worked on the weekend we calculate:

A) 6.5 x 4.75 C) 6.5 - 4.75

B) 6.5 ÷ 4.75 D) 6.5 + 4.75

---

116. Mary had \$40.00, and spent \$18.75 for a radio. To find how much money she had left, we calculate:

A) \$40.00 ÷ \$18.75 C) \$40.00 + \$18.75

B) \$40.00 x \$18.75 D) \$40.00 - \$18.75

Directions: In problems 117 to 125 solve each and choose the correct answer.

---

117. What is the cost of 2.8 pounds of meat selling for \$1.35 a pound?

A) \$4.15 C) \$1.45

B) \$3.78 D) \$1.63

---

118. If you saved \$16 a week, how long would you have to save to buy a television set costing \$288?

A) 12 weeks C) 272 weeks

B) 304 weeks D) 18 weeks

119. Susan had $565.20 in her savings account. She then deposited $150 into her account. How much did she have in her account?

A) $715.20
C) $566.70
B) $415.20
D) $563.70

---

120. It took 14.7 gallons of gas to fill the 21 gallon tank of a car. How much gas was in the tank before it was filled?

A) 35.7 gallons
C) 6.3 gallons
B) 7.7 gallons
D) 7.3 gallons

---

121. Susan had to pay an 8% sales tax on an $8.00 sweater. How much was the sales tax?

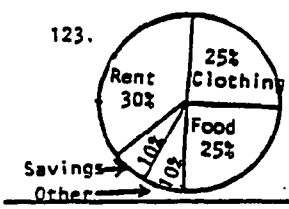
A) $.08
C) $.80
B) $.64
D) $6.40

---

122. If $40 was withheld in taxes from a $160 paycheck, what percent of the total was withheld in taxes?

A) 40%
C) 50%
B) 25%
D) 8%

---

123. 

The "circle" graph shows a budget for $400. How much money is spent for food?

A) $10
C) $80
B) $25
D) $100

---

124. Tom bought a new tennis racket for $6.00 that originally cost $8.00. What percent was the tennis racket reduced?

A) 25%
C) 14%
B) 20%
D) 2%

---

125. A camera is regularly priced at $40. During a sale it was sold at 20% off the regular price. What was the amount of discount?

A) $8
C) $32
B) $20
D) $48

APPENDIX II

ITEMS ELIMINATED IN RESCORING TEST 2

Items Eliminated in Rescoring Test 2:
Items Identified as Biased by LOGIST for
Either Ethnic Subgroups, Gender Subgroups,
or Both Ethnic and Gender Subgroups

Item Numbers				
10	38	59	76	111*
11	44*	60	80	112*
12	45*	61	82	
17	46	66	83*	
18	48	68	84	
22	51	69	90	
23	53	70	98	
25	55	72*	100	
31	56	73	103	
35	57	74	105	
36	58	75	106	

*Item identified as biased by LOGIST for both ethnic and gender subgroups.

References

Alderman, D.L. and Holland, D.W. Differential item performance across native language groups on a test of English proficiency. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles, April 1981.

Angoff, W.H. The investigation of test bias in the absence of an outside criterion. Paper presented at the National Institute of Education Conference on Test Bias, Annapolis, December 1975.

Angoff, W.H. The use of difficulty and discrimination indices in the identification of biased test items. Paper presented at the Third Annual Johns Hopkins University National Symposium on Educational Research, Washington, D.C., November 7, 1980.

Angoff, W.H. and Ford, S.F. Item-race interaction on a test of scholastic aptitude. Journal of Educational Measurement, 1974, 34, 807-816.

Berk, R. Introduction to the Third Annual Johns Hopkins University National Symposium on Educational Research. Paper presented at the Third Annual Johns Hopkins University National Symposium on Educational Research, Washington, D.C., November 7, 1980.

Bianchini, J.C. Achievement tests and differential norms. In M.J. Wargo and D.R. Green, Achievement Testing of Disadvantaged and Minority Students for Educational Program Evaluation. New York: CTB/McGraw Hill, 1977.

Birnbaum, A. Test scores, sufficient statistics and the information structures of tests. In F.M. Lord and M.R. Novick, Statistical Theories of Mental Test Scores. Reading, Mass.:Addison-Wesley, 1968.

Burril, L. Empirical comparisons of item bias methods. Paper presented at the Third Annual Johns Hopkins University National Symposium on Educational Research, Washington, D.C., November 7, 1981.

Cleary, T.A. Test bias prediction of grades of Negro and white students in integrated colleges. Journal of Educational Measurement, Summer, 1968, 5, 115-124.

Cole, N. Comment of judgmental methods for item bias detection. Comments made at the Third Annual Johns Hopkins University National Symposium on Educational Research, Washington, D.C., November 7, 1980.

Cohen, J. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin, 1978, 70, 213-220.

Coffman, W. Bias methods used by test publishers. Paper presented at the Third Annual Johns Hopkins University National Symposium on Educational Research, Washington, D.C., November 7, 1980.

Cooley, W. and Leinhardt, G. Design and educational findings of the instructional dimensions study. Paper presented at the annual meeting of the American Educational Research Association, Toronto, March 1978.

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

Cronbach, L. J. Five decades of controversy over mental testing. American Psychologist, 1975, 20, 1-14.

Cronbach, L.J. Validity on parole: How can we go straight? New Directions for Testing and Measurement, 1980, No.5, 99-108.

Diana et al. v. California State Board of Education. U.S. District Court for the Northern District of California (consent decree), 1970.

Darlington, R.B. Another look at "culture fairness." Journal of Educational Measurement, 1971, 8, 71-82.

Davies, M. and Fleiss, J.L. Measuring agreement for multinomial data. Paper presented at the Biostatistics Seminar at the Columbia School of Public Health, 1982.

Fishbein, A.L. An investigation of the items in a test battery. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, D.C., April 1975.

Fleiss, J.L. and Cohen, J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educational and Psychological Measurement, 1973, 33, 613-619.

Freeman, D.J., Kuhs, T., Porter, A.C., Knappen, L.B., Floden, R.B., Schmidt, W.H., and Schwille, J.R. The fourth grade mathematics curriculum as inferred from textbooks and tests. Paper presented at the annual meeting of the American Educational Research Association, Boston, April 1980.

Green, D.R. Procedures used for reducing bias in tests at CTB/McGraw Hill, 1966-1980. Paper presented at the Third Annual Johns Hopkins University National Symposium on Educational Research, Washington, D.C., November 7, 1980.

Green, D.R. and Draper, J.F. Exploratory studies of bias in achievement tests. Paper presented at the annual meeting of the American Psychological Association, Honolulu, September 1972.

Griggs et al. v. Duke Power Company. U.S. Supreme Court, No. 124, October Term, 1970.

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

Hambleton, R.K. Review methods of criterion-referenced test items. Paper presented at the annual meeting of the American Educational Research Association, Boston, April 1980.

Hambleton, R.K., Swaminathan, H., Cook, L., Eignor, D., and Gifford, J. Developments in latent trait theory: Models, technical issues and applications. Review of Educational Research, 1978, 118, 467-510.

Hobson v. Hansen. U.S. District Court for the District of Columbia, 269F., Supp. 401, 1967.

Hunter, R.V. and Slaughter, D.D. ETS Test Sensitivity Review Process. Princeton, NJ: Educational Testing Service, July 1980.

Ironson, G.H. and Subkoviack, M.J. A comparison of several methods of assessing item bias. Journal of Educational Measurement, 1979, 16, 209-215.

Jensen, A. Bias in Mental Testing. New York: The Free Press, 1980.

Kifer, E. and Brumble, W. The calibration of a criterion-referenced test. Paper presented at the annual meeting of the American Educational Research Association, April, 1974.

Kuhs, T., Schmidt, W., Porter, A., Holden, R., Freeman, D., and Schwille, J. A taxonomy for classifying elementary school mathematics content. Research series No. 4, East Lansing, Michigan: Institute for Research on Teaching, Michigan State University, April 1979.

Larry P. et al. v. Riles. U.S. District Court for the Northern District of California, 1972.

Leinhardt, G. and Seewald, A.M., Overlap: What's tested, what's taught? Paper presented at the annual meeting of the American Educational Research Association, Boston, M.A. 1980.

Lord, F. Applications of Item Response Theory to Practical Testing Problems. Hillsdale, NJ: Lawrence Erlbaum Associates, 1980.

McCarthy, K. Sex bias in tests of mathematical aptitude, Ph.D. dissertation, City University of New York, 1975.

Medley, D.M. and Quirk, T.J. The application of a factual design to the study of culture bias in general culture items on the National Teacher Examination. Journal of Educational Measurement, 1974, 11, 235-245.

Merz, W.R. Test fairness and test bias: a review of procedures. In M.J. Wargo and D.R. Green, Achievement Testing of Disadvantaged and Minority Students for Educational Program Evaluation. New York: CTB/McGraw-Hill, 1977.

Merz, W.R., and Grossen, N.E. An empirical investigation of six methods for examining test item bias. California State University, Sacramento, Calif., 1978.

Merz, W., Grossen, N., Groome, M., Groome, W. An empirical investigation of six methods for examining test item bias: Final Report. California State University, Sacramento, California, 1979.

Merz, W.R. and Rudner, L.M. Bias in testing: a presentation of selected methods. Paper presented at the annual meeting of the American Educational Research Association and the National Council for Measurement in Education, Toronto, March 1978.

Nungster, R.J. An empirical investigation of three models of item bias. (Doctoral dissertation, The Florida State University, 1977). Dissertation Abstracts International, 1977, 38, 2726A.

Peterson, N.S. and Novick, M.R. An evaluation of some models for culture-fair selection. Journal of Educational Measurement, 1976, 13, 3-29.

Porter, A.C., Schmidt, W.H., Floten, R.E., Freeman, D.J., Impact on What?: The Importance of Content Covered. East Lansing, Michigan: Michigan State University, Institute for Research on Teaching, 1978a (Research Series No. 2).

Porter, A.C., Schmidt, W.H., Floden, R.E., and Freeman, D.J. Practical significance in program evaluation. American Educational Research Journal, 1978b, 15 (4), 529-539.

Potthoff, R.F. Statistical aspects of the problems of biases in psychological tests. University of North Carolina Institute of Statistics Mimeo Services, No.479, 1966.

Qualls, A. and Hoover, H. Black and white teacher ratings of elementary achievement test items for potential race favoritism. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles, April 1981.

Raju, N.A. A review of item bias detection procedures employed with the SRA achievement series. Paper presented at the Third Annual Johns Hopkins Symposium on Educational Research, Washington, D.C., November 7, 1980.

Reckase, M.D. Ability estimates and item calibration using the one and three parameter logistic models: A comparative study. Technical Research Report 77-1, Personnel and Training Research Programs, Office of Naval Research, October, 1977.

Rosenfeld, M. and Thornton, R.F. The development and validation of a police selection examination for the city of Philadelphia. Princeton, New Jersey: Educational Service Center for Occupational and Professional Assessment, 1974.

Rudner, L.M. An approach to biased item identification using latent trait measurement theory. Paper presented at the annual meeting of the American Research Association, New York, April 1977.

Rudner, L.M., Getson, P.R., and Knight, D.L. The effects of various tests and item properties on five approaches to biased item detection. Paper presented at the meeting of the National Council on Measurement in Education, San Francisco, April 1979.

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

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

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

Scheuneman, J.D. A posteriori analyses of biased items. Paper presented at the Third Annual Johns Hopkins University Symposium of Educational Research, Washington, D.C., 1980.

Schmeiser, C.B. An approach to the study of item bias using experimental design. Paper presented at the Third Annual Johns Hopkins University Symposium of Educational Research, Washington, D.C., 1980.

Shepard, L.A. Definition of bias. Paper presented at the Third Annual Johns Hopkins University Symposium on Educational Research, Washington, D.C., 1980.

Shepard, L., Camilli, G. and Availl, M. Comparison of six procedures for detecting test item bias using both internal and external validity criterion. Paper presented at the annual meeting of the National Council on Measurement in Education. Boston, April 1980.

Spenser, T.L. An investigation of the National Teacher Examination for bias with respect to black candidates. Dissertation Abstracts International, 1973, 33 (8).

Thorndike, R.L., (Ed.). Educational Measurement (Second Edition). Washington, D.C. American Council on Education, 1971.

Tittle, C.K. Fairness in educational achievement testing. Education and Urban Society, 1975, 8, 86-103.

Tittle, C.K. Judgmental methods in test development. Paper presented at the Third Annual Johns Hopkins University National Symposium on Educational Research, Washington, D.C., 1980.

Wood, R.L. and Lord, F.M. A user's guide to LOGIST. Research Memorandum. Princeton, New Jersey: Educational Testing Service, 1976.

Wood, R.L., Wingersky, M.S. and Lord, F.M. LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters. Research Memorandum, Princeton, New Jersey: Educational Testing Service, 1976.

Wright, B., Mead, R. and Draba, T. Detecting and correcting test item bias with a logistic response model. Research Memorandum. Number 22. Statistical Laboratory, The University of Chicago, October 1976.

Wright, B.D. and Stone, M.H. Best Test Design. Chicago: Mesa Press, 1979.