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

Stein, Susan Volpert

AN INVESTIGATION OF ITEM CHARACTERISTICS WHICH ARE PREDICTIVE OF ITEM BIAS

City University of New York PH.D. 1984

AN INVESTIGATION OF ITEM CHARACTERISTICS WHICH ARE PREDICTIVE OF ITEM BIAS

by

SUSAN VOLPERT STEIN

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.

4/11/84
Date

Chairman of Examining Committee

4/11/84
Date

Executive Officer

Dr. David Rindskopf

Dr. Max Weiner
Supervisory Committee

The City University of New York

Abstract

An Investigation of Item Characteristics Which Are Predictive of Item Bias

By

Susan V. Stein

Advisor: Professor Alan Gross

The primary purpose of this study was to determine if a set of observable item characteristics could be identified a priori which would be predictive of a statistical index of item bias. First, data was analyzed from the standardization sample of the Metropolitan Achievement Test, 1978, Form J (MAT-J) for grades seven, eight and nine to establish the criteria by which the item characteristics would be scaled on the Reading, Math and Science subtests. A set of seven item properties were defined for each subtest. The full chi-square index was used as the measure of item bias and developed for each item as both a continuous and dichotomous value. In addition, four different methods for constructing the majority (i.e. White) and minority (i.e. Nonwhite) groups were employed. Then, prediction equations were derived based upon the multiple correlation of these characteristics with the eight chi-square indices. Cross-validation of these prediction equations were done on Form K of the MAT, 1978 (MAT-K) with a different sample of examinees of the same age, and on the General

Educational Development Test (GED) a high school equivalency test for examinees aged 17 and older.

The results indicate that item characteristics can be defined which correlate with the measures of item bias. It was found that the prediction equations worked better for the MAT-K, a different form of the same test with the same age group than for the GED, an entirely different test administered to a different age group. The practical implications of the a priori identification of biased items are enumerated as well as suggestions for future research.

Acknowledgements

To my mother, father and brother who motivated my interest in knowledge.

To my husband Steve who helped, inspired and has always encouraged that interest throughout the years.

To my children, Matt, Danny and Victoria who daily demonstrate how incomplete is that knowledge.

And to Alan Gross whose patience, guidance and expertise were essential to the completion of this dissertation.

TABLE OF CONTENTS

Chapter		Page
I :	Introduction.	1
II :	Review of Related Research.	11
III:	Statement of Problem.	83
IV :	Methodology	87
V :	Results	120
VI :	Discussion and Recommendations. .	142
VII:	Bibliography.	158

List of Tables

		Page
Table 1	Ethnic composition of examinees.	93
Table 2	Number of examinees in each sampling procedure	102
Table 3	Selection of common set of seven characteristics for Math . .	110
Table 4	Means, standard deviations and group sizes of test scores	121
Table 5	Proportions and frequencies of significant chi-square values. . .	123
Table 6	Means and standard deviations of independent variables	127
Table 7	Raw regression coefficients for MAT-J.	130
Table 8	Standardized regression coefficients for MAT-J	131
Table 9	Multiple correlations of chi-squares with item properties for the MAT-J.	137
Table 10	Correlation of predicted chi-square with observed chi-square by subtest	139

List of Figures

		Page
Figure 1	Dependent Variables. . . .	105
Figure 2	Means of continuous chi-squares for each subtest.	124

CHAPTER I: INTRODUCTION

The detection of biased items within a test has been a major concern of educational measurement for the past decade. At issue has been what constitutes bias in an item and how the test developer can eliminate, or minimize item bias from a test. Item bias is defined in general as a differential response based upon group membership (e.g., sex, ethnicity) rather than the ability the test purports to measure. In other words, the probability of success on an item is not simply a function of examinee ability, but rather could be estimated from the examinee's group membership. Item bias is a relative term: the procedures presently used select an item as biased in comparison to the other items within that particular test.

Most work to date has focused on the techniques or strategies which can identify an item as biased, but little attention has been given to the next step: once items are identified as biased, what revisions, if any, can be made to these items to reduce the bias within the test. More importantly, can observable features of biased items be identified so that item writers can use this information a priori? The purpose of this dissertation was to investigate whether certain observable item properties could be isolated which would

predict the degree to which an item would be classified as biased.

Previous research in this field can be categorized broadly into two areas: (1) development of statistical procedures to detect biased items within a test *and* (2) investigation of the empirical item characteristics which may be related to the level of item bias. In the first instance, many procedures have been developed to identify items for which differential group response patterns exist. Comparisons of transformations of each group's item p-values, differences between point-biserial indices for the groups, and differences in the factor structures of each group have been examined in various ways to identify biased items. In addition, latent trait models, which relate the probability of success on an item to some underlying trait or ability, have been used in classifying biased items. In these models the measure of bias reflects the degree to which this probabilistic relationship changes across groups of examinees. These different statistical methods do not always yield the same results.

The research into the development of statistical procedures for identifying biased items is basically post hoc in nature. The major question unanswered by these methods is what are the empirical

characteristics of biased items. In other words, given an item with certain observable characteristics what is the probability of that item being biased? In general, the research which has addressed this question has been inconclusive. For instance, several studies have attempted to examine biased items to determine what observable features or properties these items have had in common. An example of this type of study was conducted by Linn, Levine, Hastings and Wardrop (1980). They proposed after identifying biased items: "to investigate the possibility that certain features of the items . . . may lead to biased estimates of the reading achievement level for black students . . . the item so identified could be characterized by some generalizable features . . . (p. 3)." The authors describe in detail the statistical procedures for the detection of biased items, but do not specify what, if any, observable features of the items will be analyzed. When the results of the study are presented and the items selected as biased are compared with those items not classified as biased, the authors comment that no "differences" were noted between the two groups of items. Typical of this type of study no operational definition of "differences" is given.

Other studies have also considered the empirical characteristics of biased items. For example, in studies by Burrill (1981), Ironson and Subkoviak (1979), Merz and Grossen (1979), Nungester (1977), Rudner, Getson and Knight (1980), and Shepard, Camilli and Averill (1980) items are first selected as biased in terms of several statistical procedures and are then inspected to identify the basis for their selection. Item type has been suggested as a possible important characteristic. For instance, Shepard, et al. state that verbal analogies appear biased against black students. However, it is difficult to generalize from these results since only some verbal analogy items were selected as biased against black examinees, while other analogy items were not selected.

One empirical characteristic which has received considerable attention is that of item content. McCarthy (1976) found that when the context of mathematical problems was familiar to males, male students scored higher than females. Also, the reverse was true when the context was more familiar to females. In a study by Medley and Quirk (1974), the inclusion of black history items improved the scores of blacks on the National Teacher Examination. However, other studies present conflicting results perhaps due to an imprecise

definition of the term "item content." Should content be defined as the reference to a particular group or group experience, such as reading passages in black history or the suffragette movement? Or, does it mean information which would be more relevant to one group than the other because of exposure or experience, i.e. cooking or football. It is difficult to determine which definition has been applied when content has been examined, but most studies appear to investigate the interest or familiarity of material to minority groups.

Since the goal of bias research is to reduce differential responding according to group membership, operational definitions of what is being assessed seem essential. The introduction of any content which would enhance one group's probability of success would seem suspect. A more reasonable approach would be to include questions and materials which are as neutral as possible and to rely on only those experiences relevant to the ability one is trying to measure.

In summary, studies to date have not systematically investigated which item features may be related to item bias. Also precise definitions of these features have been lacking. The previous work has focused on reviewing items selected as biased in a specific test in order to discover if these items

possessed a dimension in common. Since most large-scale test development proceeds through a complex carefully-monitored process, it would seem reasonable that more than one or two features of the item could be responsible for its selection as biased and that more than one inspection of biased items would be needed to verify this hypothesis. Rather, much of the variance in a statistical item bias technique might possibly be accounted for in terms of some combination of measurable and observable item characteristics.

This study analyzed the Metropolitan Achievement Test (MAT), Form J, 1978 Advanced 1 for grades 7, 8 and 9. All seventh and eighth grade examinees who had completed at least half of each subtest and on whom there was biographical information (i.e. ethnicity and sex) were included. The examinees were divided into two groups, Whites and Others, i.e. nonwhites, and from these two groups four sampling procedures were developed. Limitations on sample size prevented the investigation of other groupings. The Reading, Math and Science subtest were used to define observable item characteristics. On each subtest a set of seven item properties were chosen on the basis of pretesting and stepwise multiple regression analyses. For the Reading

subtest these included: 1) Reference to any group 2) Similarity to practice item 3) Length of item stem 4) Difference in the length of the response options 5) Question format 6) Item location and 7) Passage length. The item characteristics which were defined for each item on the Math subtest were: 1) Similarity to practice item 2) Length of item stem 3) Difference in length of response options 4) Question format 5) Response format 6) Arrangement of response options and 7) Type of problem. On the Science subtest each item was scaled on the following properties: 1) Length of item stem 2) Difference in length of response options 3) Arrangement of response options 4) Total number of words in response alternatives 5) Number of words in correct answer 6) Item location and 7) Presence of visual material. These characteristics were scaled for each item on the appropriate subtest and the results were used to predict the degree of bias for each item.

The full chi-square method (Shepard, Camilli & Averill, 1980) was employed to assign a statistical bias index to each item. This technique computes contingency tables for each item which includes both correct and incorrect responses for each group. To control for differences in ability between the two groups each item is divided into discrete score intervals based

upon the individual's total score. Within each score interval an unbiased item is defined as one where the probability of a correct response is the same for all examinees of that ability irrespective of group membership. The chi-square value for each item was a summation of the chi-square values for each score interval and represents bias against all groups. A signed index was also calculated for the Matched sample. This Matched sample consisted of equal numbers of Whites and Others, where the Whites were selected from the Total group to have the same total score distribution as the total Other group. This index attached a positive or negative sign to the chi-square at each score level based on whether the observed number of correct answers was the same as the expected number. If the observed frequency was less than the expected frequency for Others a positive sign was given the chi-square, indicating bias against Others. If the observed frequency of correct answers was less than the expected frequency for Whites, a negative signed was attched. In this way, bias against Whites was subtracted from the index and only bias against Others remained. The actual chi-square value for each item was treated as the continuous dependent variable and also a dichotomous variable was established between a biased item and an unbiased one.

The major analysis of this dissertation used multiple regression to see the degree to which the independent variables, or item characteristics, predicted the dependent variable, a measure of item bias. Form J of the MAT was used to define item characteristics for each subtest and to develop prediction equations. Then, these prediction equations were cross-validated on the Reading, Math and Science subtests of the MAT-Form K, a different form of the same test with a different sample of examinees and the General Educational Development (GED) Test, a high school equivalency test administered to examinees 17 and older. It was hoped that a prediction scheme could be developed from the initial data on the MAT-J which then could be validated on the MAT-K and GED.

Investigation of the hypothesis that there is a positive relationship between some combination of observable item features and an index of item bias has important implications. While the search for the single best and/or simplest strategy for detecting bias should continue, that research is ex post facto and does not help test developers at the outset. The main issue is to develop testing instruments which are as bias-free as possible and to aid in that development by furnishing empirical evidence for item writers and test publishers

which can be used *a priori*. Content specifications are of primary importance in test development, but attention should be given to other aspects of an item which can influence an examinee's response. By focusing on item properties which can be defined operationally and analyzed objectively, the total effect of these properties can be measured, as well as the contribution of each item feature. The results of this study should provide some specific guidelines for item writers.

CHAPTER II: REVIEW OF RELATED RESEARCH

The investigation of test bias can be divided into two categories: (1) bias in the fair use of the test and (2) bias in the test itself. Generally, the first type of study looks at the use of the test score in selection procedures (Cleary, 1968; Cole, 1968; Darlington, 1971; Einhorn & Bass, 1971; Flaugher, 1978; Gross & Su, 1975; Linn, 1973; Peterson & Novick, 1976; Thorndike, 1971). In all these studies the definition of "fair" varies with the mathematical model presented, and fair test use is the concern, as opposed to unbiased measurement.

The present study addresses the second issue, which is bias within the test. The question answered here is: Do persons of equal ability have the same probability of obtaining the correct response to an item or can their response to an item be predicted by their group membership regardless of ability? For instance, will a male student and female student have the same probability of answering a math problem correctly, given that they possess the same degree of mathematical ability?

The problem which has confronted researchers who are investigating differential response patterns

based upon group membership can be stated as follows: Define a procedure which analyzes item statistics for different groups, such as item difficulty or item discrimination on a particular test or subtest; compare the performance of these different groups on the individual items in terms of those statistics; and select those items which appear different based upon the results of comparing those statistics. Not only is the definition of similar ability difficult, but also a decision as to how much deviation from other items can be allowed before an item is termed "biased" must be made. The determination of deviance among items and the extent of justifiable deviance between groups are the major considerations and differences between statistical procedures designed to detect bias within a test. Additional confusion arises when one method is cited but a slightly different procedure is followed so that it is difficult to compare results from supposedly similar methods as well as results across different methods.

Initially the inspection of group means was used to determine if different groups performed differently on the same item. Since this is a measure of the distribution of scores within each group as well as the ability level of the group, it was deemed an unsatisfactory strategy in selecting items on which

various groups responded differently (Jensen, 1980). Cardall & Coffman (1964) were the first to use the analysis of variance design to test for the presence of an item x group interaction. This meant that different groups performed differently on the same item; that selected items were easier for one group than another. In their procedure, group membership represented one factor and item scores a repeated measures factor. Three samples of 300 answer sheets were selected from the May 1963 administration of the SAT. The three groups were taken from rural towns in the midwest, New York City and centers from the southeast where only Blacks had registered. Analyses were done separately for the verbal and the mathematics sections. Significant main effects and item x group interactions were found on both sections.

The data were further analyzed to see if the within-group differences of item difficulty were as large as the between-group differences. Two of the samples within each group were used to assess within-group differences while the third group was used to determine the correlations between the item difficulties of the three groups. The within-group correlations were found to be higher than those between groups for verbal items, which demonstrated the lack of correspondence between

the relative item difficulties for the Black group and the other two groups. The between-group correlations for the math items were found to be similar. Most importantly, Cardall & Coffman found that the items did not have the same ranking across groups, but it could not be determined whether the bias was balanced among the groups or favored one group.

Further extension of the ANOVA method was employed by Cleary & Hilton (1968) in their study to determine the appropriateness of the PSAT for Blacks. They analyzed data from the 1961 and 1963 PSAT administrations and selected samples from seven integrated schools in three large metropolitan centers. The items were scored one point for a correct answer, zero for no response and minus one-fourth for a wrong answer. A three-factor design was used with items, race and SES within race. All main effects and interactions were found to be significant. These results are in part attributable to the large sample sizes. Estimating the variance components and percentage contribution of each effect to the total variance of a single observation chosen at random would be a better approach. The smallest contribution to the total variance were Item x Race interactions and Item x SES Within Race interactions; while the most variance was contributed

by Subject x Item interaction. (Hoyt (1941) refers to this as variance due to error of measurement.) Cleary & Hilton conclude that the PSAT is not biased for the groups studies.

The Transformed Item Difficulty (TID) approach, the most widely used procedure to detect item bias, relies on the item x group interaction definition, but is a graphic representation of each item difficulty plotted separately for each group. Originally it was discussed by Angoff in 1972, who attributed the derivation to Thurstone's work on absolute scaling (1925). Angoff & Ford (1973) used this method on random samples of Blacks and Whites on the PSAT. First, the p-values or item difficulties for each item and each group are computed. In order to linearize the relationship between these values for two groups, each p-value is transformed to a z-value corresponding to the (1-p)th percentile of the standarized normal distribution. To eliminate negative z values delta values are calculated as $\Delta = 4z + 13$. The pairs of deltas are plotted on a bivariate graph, where the delta values for one group are on the abscissa and the other group on the ordinate. Generally, if groups are similar, the plot of these points should form a long narrow ellipse which would indicate a high correlation between the scores on

the items of the groups. Conversely, the lower the correlation of the scores between groups the distribution of points and the shape of the plot will change.

Of greatest interest are those points which fall at some distance from the major axis of the plot. These points represent items which contribute to the item x group interaction and on which the different groups under investigation performed quite differently: that is, the rank order of item difficulty for the two groups is not the same. Angoff & Ford computed the major axis line of the ellipse and calculate the perpendicular distance from that line for each item. This is the measure of the deviation of each item which is used to determine whether an item is biased. However, how much deviation to allow before an item is labelled "biased" is not discussed.

Many variations of the TID approach have been used in recent studies. These variations occur in the transformations (or not) of the difficulty value, the calculation of the main axis of the ellipse and the definition of what constitutes a biased item. Coffman (1961) had devised a method for transforming the item difficulty values for two groups which plotted paired values of 2 arc sine $\sqrt{p}$. In this way the p-values have the same standard error, however, like p-values they are

bounded. Instead of the transformation to normal deviates, Merz & Grossen (1979) and Rudner, Getson & Knight (1980) calculate standard scores for each item and each group utilizing item means and standard deviations for that group. In this way the groups are adjusted to the same mean standard deviation so that the 45° line is taken as the theoretical line of no bias. Fishbein (1974) also used nontransformed p-values. The absolute difference between mean difficulty values for a pair of groups was computed and then compared to the difference between two groups on a specific item. One serious problem here is that if there is any curvilinearity in the relationship between z-values, it would be misinterpreted as bias in the item.

Echternacht (1974) transformed p-values to deltas, but then obtained the difference for each pair of deltas. These difference values were plotted on normal probability paper and the mean and variance of this distribution of differences obtained. Then, the hypothetical normal distribution was plotted with confidence bands drawn. Items outside these bands were considered biased.

Another way to define the major axis has been proposed by Sinnott (1980) whose concern was the items selected as biased. First, a decision is made as to how

much deviation may influence the computation. Then, she calculates the major axis with all the items included and removes those items which meet her definition of bias. Now, a new line is computed without those items, but she readmits those items to the new set of points to see if they still meet the definition of bias and to see if any new items would be selected. If the items are the same and no new items are found, the major axis line has been determined, if, on the other hand, new items are found, they would be removed (along with the first group) and the line recalculated. This process continues until no new items meet the definition of bias. It is possible that so many items would be specified as biased that the procedure may have to be redone with a less stringent definition.

The cutoff point between an item designated as biased and one that is not, differs from study to study. Angoff (1972) recommends drawing confidence bands on either side of the main axis line, but in his studies (Angoff & Ford, 1973; Angoff & Sharon, 1974; Angoff, 1975) he computes the perpendicular distance of an item to the line. Reference is made to "extreme" items with no definition of that term. Generally, studies which have used the delta-distance measure have defined bias as those items which are 1.5 standard deviations from the mean of the distribution (Burrill, 1981; Nungester,

1977). In their study of sex bias Strassburg-Rosenberg & Donlon (1975) define greater than 1.5 standard deviations of the residuals as items which are biased, while Rudner (1978) uses .75z value as the cut score for deviant items. An empirical investigation into cut scores was suggested by Sinnot (1980) who uses .3 as the distance beyond which to define biased items.

The major criticism of the TID method is the confounding that occurs between differences in group ability, item discrimination and item difficulty (Hunter, 1975; Lord, 1977). Unless all the items in the plot have the same discriminating power for groups of different abilities, the items selected by this method as biased may be those items which discriminate best between those of high and those of low ability. Hunter also demonstrates how an unbiased test can show item x group interaction with items of varying difficulty. Fully aware of these problems, Angoff (1975) makes two suggestions. First, use groups that have been matched on ability (Angoff & Ford, 1973). Although this appears very logical, until recently few studies have formed matched groups as a subanalysis when the TID method has been used. All those studies found fewer items to label biased (Burrill, 1981; Ironson & Subkoviak, 1979; Jensen, 1980; Rudner, 1978). These findings support the notion

that the difference in ability between groups can obscure the bias measure.

Angoff's second suggestion to researchers was to use pseudogroups (Jensen, 1980). These groups provide a measure of how much deviation to allow before an item is classified as biased. Two groups are formed from the majority group with the group menas and distributions differing in the same way as the cross-cultural groups. For example, two male groups are selected where one group's mean is the same as the male group and the other group mean is the same as the female group in the study. In this way, it is possible to determine if items were selected as biased because of the differences in group membership, or because of differences in group ability. Pseudogroups examine the acceptable spread about the axis from within-group differences which exist before the differences are attributed to between-group differences.

Traditionally, item discrimination indices, the point biserial or biserial correlation, have been used in test development to insure the reliability of the test. They are a measure of the correlation between an item and performance on the total test of which the item is part. In other words, these correlations describe the degree to which all items are measuring the same trait, whatever that trait is, for all examinees.

Since the definition of an unbiased item is an item where all examinees of the same ability have the same probability of success, that is to say an item measures the same trait equally in persons of the same ability, the use of these discrimination indices holds intuitive appeal.

Green & Draper (1972) used the point-biserial correlation in their study which analyzed data from the CAT battery with seven groups of subjects. The test was separated into the "best" half and the "worst" half on the basis of their item-test correlations. The seven groups were paired with each of the remaining six and biased items were defined as those that fell into the "best" in one group but not the other. The results showed that, as anticipated, the proportion of biased items was less for similar groups than for unlike groups. When the mean differences between the major and minor groups were inspected, it appeared that items chosen as "best" for the major group increased the difference in means between the major and minor group while items selected as "best" for the minor group reduced the difference in means. Strangely, the authors reported "biased" items, in terms of their definition, when they compared similar groups such as two white suburban high SES groups.

Hunter (1975) constructed a hypothetical test of six items of different difficulty but equal quality. It was claimed that each of the items was unbiased. He demonstrated how the point-biserial correlations would be different as a function of their difficulty values. Consequently, if each item is more difficult for one group than another, then different items would have to be selected as "best" for each group. Fifty per cent of the items would be labelled "biased" even though all six items were not.

The important question of whether a test is measuring the same ability in all groups has prompted researchers to use factor analysis (Green & Draper, 1972; Green, 1976; Merz 1973, 1976; Merz & Grossen, 1979). In achievement tests, each item or subset of items which produce subscores is treated as a variable. The analysis is done separately for each group to see if the different groups will produce a similar set of factor loadings. Different sets of factor loadings would indicate that the different groups did not respond to the items in the same way. Consequently, the item is considered biased because it appears to measure different traits across groups. Those items with the largest differences in factor loadings are considered most biased.

An inter-group factor analysis model used by Green & Draper (1972) and Green (1976) is based on work by Tucker (1958). Here item variance is partitioned among factors common to each subgroup, factors specific to the subgroups and residual or error variance. An item is unbiased if the proportion of variance accounted for, specific to the group, is small. Analyzing data from 270 Black and 360 White fifth graders who took the Reading Comprehension subtest of the 1970 CAT, the proportion of variance attributable to group-specific factors was similar. However, if a cutoff of 25% of group-specific variance was defined as determining bias, then nine of 42 items would be biased.

Another factor analytic approach developed by Merz (1973, 1976) incorporates factor scores and analysis of variance. The item intercorrelation matrix is computed for subjects pooled across groups. The matrix is reduced with principal components analysis and rotated orthogonally. An analysis of variance is conducted on each set of factor scores. Multiple group membership is the independent variable and the factor scores for each vector are dependent variables. Differences in average factor score for a specific group indicate bias. Merz & Grossen (1979) claim that since item intercorrelations are analyzed, mean total score

differences should not influence the selection of an item as biased.

Merz (1976) mentions an alternative strategy in which group membership is entered into the item intercorrelation matrix as dummy variables. The Goodenough-Harris Drawing Test, an instrument designed to measure mental ability was administered to groups of school age Anglo, Black, Mexican-American and Indian children. Several factors were identified but the performance patterns for the various factors were not uniform across groups. Again principal components analysis is used to reduce the matrix and after orthogonal rotation, factor loadings are inspected. Items are considered biased which have major loadings on the same variables as group membership. The magnitude of the correlations are affected by correct-incorrect scoring and again mean differences in total score should not influence the outcome.

Rudner, Getson & Knight (1980b) point out that, "The decision problems that beset factor analysis in general are multiplied when applying these factor analytic techniques to test bias" (p. 225). Which correlation matrix to inspect, what rotation to employ and how many factors to extract are but some of the problems a researcher must solve when factor analysis

is undertaken. Much of the data analysis required is not usually employed during an item tryout and the procedures are complex. In fact, Rudner and Convey (1978) found that items were selected as biased by these procedures in the pseudocultural group comparison, which make their meaning unclear.

A method which does not examine correct responses or use the total score of an individual, but rather looks at the pattern of incorrect choices among groups, has been developed by Veale & Forman (1976). They claim a correct answer is a correct answer for everyone, so that bias in test items must be a function of the "foils", or incorrect answers, from which an examinee selects. Bias is defined as the differential pull of the distractors for different groups. In this procedure a chi-square technique is used with columns for each wrong answer (distractor or foil) and rows for each group. The null hypothesis to be tested is that, for those examinees who got the answer wrong, the proportion who selected each response is the same irrespective of group membership. Each item is inspected to see whether there is an association between the selected distractor and group membership.

This procedure is computationally simple and theoretically easy to understand but the underlying

assumption about distractors is not true for most items. This assumption is that the foils are "equally" attractive to one who does not know the correct response. Generally, wrong answers "distract" the examinee from the correct answer in varying degrees which represent a common error or partial knowledge, so that the "pull" of each choice is not the same. Also, while ability is not measured by total score and does not enter into the computation in any way, differences in distractor distributions may be actually identifying differences in the abilities between groups.

Item response theory (IRT), also called latent trait theory, is based upon the notion of sample invariance: that is to say, the measure does not depend upon the distribution of ability of the particular samples, but that the probability of a correct response is the same for all individuals at a given ability. The probability of a correct response does not depend upon how many other examinees are located at the same, or some other, point on the scale. IRT describes a statistical model which connects an unobservable latent trait to an observable test score. The various statistical models developed from IRT specify the relationship between observable examinee test performance and traits, or

abilities, which are assumed to underlie the performance on a test.

Three important assumptions provide the foundation for IRT: (1) unidimensionality; (2) local independence; and (3) item characteristics curves (ICC). It is assumed that a single, underlying ability is being measured by the test, or subtest. Often factor analysis is used to test for unidimensionality and if more than one large factor is found, the test is divided into clusters corresponding to the factors and assessed separately. Local independence means that a response on one item is not dependent upon the response from another item, given some fixed ability level. The ICC is the graphic representation of the mathematical function which relates the probability of a correct response to an item to the ability measured by the test. It is the regression function which fits performance on an item to the ability scale. An ICC is defined completely when its general from is specified and when the parameters of the curve for a particular item are known.

The different IRT models use a different number of parameters to describe the ICC. In the one parameter model (Rasch, 1966; Wright & Stone, 1979) the ICC's are non-interacting curves which vary only by a translation

(b) along the ability scale, or, in other words, items differ only in difficulty. The ICC in the two-parameter model (Birnbaum, 1968) differ in slope (a) and translation (b). Some items are more discriminating than others with corresponding increases in the curves as well as differences in item difficulty. In the three-parameter model (Lord, 1977; Lord, 1980; Lord & Novick, 1974) items are said to vary in slope, translation and lower asymptote (c). Originally, this was specified as a guessing parameter but because estimates were often less-than-chance, it is now known as a pseudo-chance estimate and is especially relevant in the lower ability range.

In IRT approaches, an item is considered unbiased if examinees from different groups who have the same ability, have an equal probability of a correct response. In other words, the ICC's are the same for the groups studied. The determination of the ICC's is different in each model depending upon the number of parameters to be estimated.

Currently, several equating methods are used for the a and b parameters of the different groups (Lord, 1980; Marco, 1977; Shepard, Camilli & Averill, 1980). In one case separate parameter estimates are obtained for each group under investigation, then a principle axis

line is fit to a bivariate plot of b values for two groups. The group's parameters, whose b values have been put on the y-axis, are adjusted to the other group by subtracting the intercept and dividing by the slope. The same procedure is followed for the a parameter except that the intercept is forced through the origin. Lord (1980) suggests a modification of the procedure recommended by Marco (1977) for test equating. Standardize on the b's, rather than on the ability estimates (θ) so that mean of the b's is 0 and the standard deviation is one. In this way, all the parameters for all the groups are on the same scale, rather than if θ were estimated first, which would possibly give each group a different mean and variance.

Once the ICC's for each group are equated for scale and plotted together, various measures are used to determine how much deviance to allow between them before an item is called biased. Rudner (1977) computed the area between the ICC's using successive rectangles with a width of .005. Lord (1980) uses an asymptotic significance test based upon the summed variance-co-variance matrices of the a and b parameters to test for significance between pairs of equated ICC's.

In the one parameter model, or Rasch model as it is also known, the index of bias, after difficulty

parameters are equated to the same scale, is the difference between the difficulty values for the two groups (Draba, 1977). However, if the data do not fit the model assumptions of constant discrimination and guessing parameters, the results will be erroneous. To overcome this problem, goodness-of-fit statistics which test the fit of the model for the groups have been devised by Durovic (1975) and Wright, Mead & Draba (1976). After the residuals have been standardized, the mean square residual for each group is calculated. This residual is the examinee's observed response minus the probability of a correct response given the person's ability estimate. An item is defined as biased if the mean square fit of the observed item is greater than one. Supposedly, bias will be detected in an item if there are differences in guessing behavior or item discrimination between the two groups, as well as differences in difficulty.

The appeal of the IRT models is sample invariance. Once the ability estimate has been defined, any sample can be put on the ability scale regardless of the ability distributions. Also, non-uniform patterns of bias can be identified because the ICC specifies the probability of a correct response at each ability. This can be seen when ICC's cross each other, which

demonstrates an opposite direction of bias at different ability levels.

Several problems remain to be solved with the use of the IRT models. By far the most serious one concerns the large sample size needed to produce stable parameter estimates. If the simplest estimation procedures are desired and the one-parameter model selected, is information lost by not including item discrimination and a pseudo-chance parameter? Conversely, analyses contingent upon the estimation of four parameters (including ability) for each item may contain magnified errors. Obtaining convergence for the estimates is difficult especially for the pseudo-chance (c) parameter. More commonly, an average c value is estimated and fixed for each item. Even when the requirements for large sample sizes and many items are met, reseachers have found that many items could not be included in the analysis because of convergence problems (Camilli, Shepard & Averill, 1980). Additionally, if the assumption of unidimensionality is not met, differences across groups caused by multidimensionality would appear as bias.

To avoid the estimation problems which occur mainly because of the extensive data requirements of the IRT models, chi-square procedures have been developed

(Alderman & Holland, 1980; Scheuneman, 1975, 1976, 1978, 1979, 1981; Shepard, Camilli & Averill, 1980). These techniques are conceptually similar to the IRT models in that an unbiased item is defined as one where the probability of a correct response is the same for all examinees of a given ability irrespective of group membership. An important difference, however, is that ability here is not sample invariant as in the IRT models, but is the total score on a homogeneous test or subtest. Generally three to five intervals are determined unless sample size is large, where more categories are appropriate (Rudner, 1977; Scheuneman, 1976). A contingency table is derived for each item where the rows represent each group (k) and the columns are divided into score categories (j). In the chi-square procedures, discrimination can vary among the items and the lower asymptote does not have to be zero. The probability of a correct answer within one score level is assumed to be constant. Instead of the smooth curve of the IRT models, the resultant probability function of the chi-square method can be thought of as rectangular steps each of which represents the height of one ability interval.

In 1975, Scheuneman published an article which describes the genesis of her procedure. Her concern was

bias in large experimental item pools and the need to evaluate specific items. Although at that time it was called a chi-square, she points out that it differs from that index, in the traditional sense, in the computation of the expected frequency of a correct response and the degrees of freedom used. Although her later articles (1976, 1978, 1979, 1980) refer to a modified chi-square, she now prefers the method to be known as C2 (1981).

In the Scheuneman method, which treats only correct answers, the score intervals are determined first. These are based upon the distribution of total scores, the need for ten to twenty observed correct responses per ability interval per group and some minimum number of incorrect responses per score interval regardless of group. Often the requirement for adequate observations in the lowest and highest ability groups are the hardest to meet. Since only correct responses are considered, Scheuneman's method can set the high ability category readily and has no difficulty setting intervals for easy items. After score intervals have been determined the next step is to calculate the expected frequency of a correct response. In each score group the proportion of correct responses is computed by dividing the total number of correct responses for all groups, by the total number of examinees within that

category. Then, this proportion is multiplied by the number in each group in that score interval in order to obtain the expected frequency. The observed frequencies are compared to the expected frequencies summed across intervals to determine the size of the C2. Large values denote bias in that item. A significance test is employed to measure the degree of group differences with (k-1)(j-1) degrees of freedom, where k equals the number of groups and j equals the number of score categories.

The Camilli chi-square (1980) makes use of both the correct and incorrect responses so that a chi-square distribution results. He analyzed data first using proportions correct across abilities for different groups then using incorrect proportions and found that the results were different. Since proportions right and wrong reflect the same information, he decided both sets of answers need to be included so that information is not lost. Actually, this procedure adds another contingency table for incorrect responses to Scheuneman's method and then the full chi-square is the sum of the chi-squares from the correct and incorrect answers. The degrees of freedom from testing this statistic are j(k-1) where j equals the score intervals and k equals the groups (Mellenburgh, 1980).

Scheuneman (1975) described four types of items which could be determined by her method in analyzing two groups: an unbiased item, an item biased toward one group, an item biased toward the other group and an item which displays a differential validity pattern. This last item is one where the bias is different for different groups at different ability levels. In other words, the item could be biased towards low ability females and high ability males. This can be compared to the ICC's crossing.

Baker (1981) sharply criticized the Scheuneman method. He demonstrates that unequal sizes of the groups will produce a spurious expected frequency of correct responses. Also, "the pooled proportion of correct responses is always intermediate to the two observed proportions" (p. 60). In response, Scheuneman (1981) counters that large differences in performance should be more important for the smaller group, who are generally the minority group. She suggests that one way to minimize the discrepancy in size between groups under study is to randomly select a subsample of the larger group to equal the size of the smaller group. She points out that the C2 approach substitutes for other, better analyses, namely IRT models, when the sample size of the groups or number of items available does not meet the

requirements for those analyses. If the incorrect answers are included in the computation more subjects, at least 200, would also be needed.

In a review of the chi-square procedures, Ironson (1982) discusses the advantages and disadvantages of the method in general and the Scheuneman and Camilli approaches in particular. While the Scheuneman technique is not distributed as a chi-square and the sampling distribution is unknown, it is useful with groups as small as one hundred. Still, Camilli's method which is distributed as a chi-square needs only 200 per group which is many fewer than the 444 to 2,137 subjects required for IRT analysis (Linn, Levine, Hastings & Wardrop, 1980). Both procedures are less sensitive to the shape of the distributions than TID, but are sensitive to the distribution of the total test score. This influences the determination of score intervals, which are arbitrary cut points. An important advantage of both is the availability of significance tests. Also, these procedures can be described easily to test takers as well as test developers.

A recent method to detect bias in an item within a test has been tried by Stricker (1981) who used the item's partial correlation with subgroup membership, total score held constant. This procedure controls for

group differences in overall ability, which is defined as the total score with the item under analysis removed. The general difficulty of the item for each group is compared and items are selected with consistent bias toward one group. This means that items which display differential validity will not be identified. Also, the partial correlation indices from different items cannot be compared directly. The index is based upon correlations for a dichotomous item which is limited to the proportion passing and failing. However, the significance of the index can be compared among items.

Stricker's purpose was to test the efficacy of his procedure compared to the TID approach and the three parameter IRT method in identifying race and sex differences on the GRE Aptitude Test. Four basic samples of white males (N=1,122), white females (N=1,471), black males (N=284), and black females (N=626) were drawn from the 1977-1978 administrations of the GRE, with corresponding samples obtained from the 1979-1980 administrations of the test to use for replication. The white samples were randomly selected from the white examinees while the black samples comprise all the qualified black examinees. The samples were limited at the outset to those examinees who did not have unusual educational background or test-taking experiences.

Supplemental analyses used special samples: several sets of samples were matched on test scores and other sets were composed of pseudogroups. These groups were used to verify the significance tests for the partial correlation method and three parameter IRT method.

The statistical analyses were computed separately for sex and racial differences. Both the partial correlation and IRT methods identified many more items of the GRE as biased than previous research investigating bias on the GRE. However, extreme outliers were not found by either partial correlation or IRT. There was considerable agreement in the number of biased items between the partial correlation procedure and the three parameter IRT in the basic analyses and in the replication except for sex differences for blacks. This may be attributed to the small sample of blacks. The agreement between partial correlation and IRT on the selection of particular items was substantial for sex differences but poor for racial differences. A possible explanation is the sensitivity of the three parameter model to differential validity, that is, items which favor one group at one ability level and favor another group at another ability level. Little consensus on individual items were found between the TID method and the other two, most likely because of the discrepancy

in the number of items chosen. Very few items were identified by the TID approach and the measure of the extent of the bias in this procedure, which was the perpendicular distance from the main axis was generally small. The items which are identified by the TID method as biased clustered near the cut score.

The use of partial correlation method in identifying items as biased appears promising. It can be used with small samples, or a number of items, and controls for differences in the ability levels of the groups being investigated. The most serious disadvantage, however, is the lack of an interaction effect. Items which favor one group at one ability level and another group at another ability level cannot be identified.

Studies which compare the results of more than one bias technique have used real and simulated data (Burrill, 1981; Ironson and Subkoviak, 1979; Merz and Grossen, 1979; Nungester, 1977; Rudner and Convey, 1978; Rudner, Getson and Knight, 1980; Shepard, Camilli and Averill, 1980; Stricker, 1981; Subkoviak, Mack and Ironson, 1981). These studies were undertaken in order to determine if the various methods were measuring the same thing and if so to determine a method of choice. One objective was to compare items chosen by the different methods to see if the same items were

specified. If the same items were selected then the "best" method would be the one which could be explained simply and could be calculated easily. However, if each method identified different items then an intercorrelation matrix was computed. If the correlations between procedures were high, it could be reasoned that these methods were detecting the same phenomena and could be used interchangeably. If different items were selected by the different methods, then it would be necessary to explore what each technique was, in fact, measuring. And finally, the items chosen by any method were to be examined in order to see if any generalizations could be made about the item content and item type of these items.

In a dissertation study completed in 1977, Ronald Nungester compared the Angoff TID method, Fishbein difficulty difference approach and Scheuneman modified chi-square procedure in order to detect sex bias in the 1970 Florida Statewide Ninth Grade Tests of English and Math. He wanted to investigate the relationships among rank ordering of items by amount of bias; the statistical and content characteristics of items identified by the three techniques; and the effects on test reliability and validity for each sex if the most biased items selected by each technique were deleted. Standard items

analysis was also used to remove the five "worst" items according to test development procedures to examine its influence on reliability and predictive validity.

Nungester found that there was a strong relationship among rank order of items identified by each technique. In fact, across both subtests at least three items were selected by two procedures and in one case, although the rank order did not agree, the Angoff & Scheuneman methods identified the same five items. Statistically, the selected items varied by subtest and technique. However, there was a tendency for biased items to be somewhat more difficult than the average even though items were chosen from throughout the difficulty distribution. Also all discrimination levels were identified but the trend was toward items with lower discrimination values.

In order to see if the removal of the five most biased items improved the reliability of the test, the author used coefficient alpha to compute the new reliability. A slight decrease in reliability was noted. It is possible that five items were not enough to remove if many more items were classified as biased. Also no correction was made for the reduced length of the test.

The predictive validity of the ninth grade test was determined with scores from the Florida Statewide Twelfth Grade Test. Using matched samples from the total population, correlations were obtained from the whole test and the subtests, as well as for the subtests with the five most biased items deleted. On the English test predictive validity improved for both sexes when the five most biased items were deleted. However, the items removed by standard item analysis improved predictive validity the most. The math results varied by sex, but no decrease in predictive validity was found. For males, improvement occurred when the five items chosen by any method were deleted but again most improvement took place when the poor items from standard item analysis were not included. For females, an increase in predictive validity occurred when items identified by Fishbein's difficulty difference technique or standard item analysis were removed.

The results of analyzing the content for sex bias were inconclusive. Nungester cautioned against making casual inferences based upon his findings. For instance some items which make reference to males were biased against males while other items with male reference were biased against females. One concludes that if standard item analysis were done separately on

samples as large as in this study, the test results would be more valid.

Rudner & Convey (1978) analyzed data from the Stanford Achievement Test, Primary Reading Comprehension that had been administered to the hearing and hearing-impaired. They compared the TID, modified chi-square, three parameter IRT, and factor score methods for identification of biased test items. Not only were they interested in whether the different approaches would select the same items, but also if these approaches would also classify item as biased in subsamples of a single population where no bias should exist.

This study exemplifies the within method differences which exist when one tries to compare results across studies which purport to use the same methodology. Rudner & Convey state they will use the TID method attributable to Angoff, but follow a different procedure. The 45° line is taken as the point of deviation instead of the major axis of the plotted points of each item. Consequently, the perpendicular distance from the major axis is not computed as the index of bias but rather items are considered deviant which are greater than a fixed item-regression line distance of .75 z-score units. To add greater confusion, a figure is described which appears to have been calculated using the Angoff method.

The setting of score intervals proved difficult in the chi-square method because there were different total score distributions for the groups so that four intervals were found to be sufficient rather than five. In addition to the usual chi-square value to indicate aberrance in the modified chi-square approach, one minus the probability associated with the chi-square was used. The problems of the chi-square method that Baker (1981) subsequently elaborated are discussed in this study. That is the spurious results obtained when the proportion of examinees passing an item is the same at each score interval but the actual numbers are very different. For example, if in score interval 2, three out of thirty in one group have the correct response and 30 out of 300 in the second group have the correct response, the proportion passing is the same. Equal score intevals are suggested to avoid this situation. However, this is hard to achieve at the highest and lowest scores because it is necessary to have both correct and incorrect responses in each score interval.

The chi-square index correlated the highest with both the TID and three parameter IRT, .59 and .67. The factor score and chi-square (1-p) methods showed the least similarity to the other techniques. The authors

concluded that these two methods are "inadequate in identifying biased items" (p. 23).

Two subsamples of the hearing-impaired group were selected so that the mean of one equalled the total hearing group and the mean of the other group was the same as the hearing-impaired. If items were identified as aberrant by any method in these groups it could be stated that the method was selecting items because of ability differences, not group differences. The TID and chi-square approaches did not identify any items in the same culture groups, whereas the IRT method identified two items not previously selected as biased. Again, the factor score and chi-square (1-p) methods demonstrated the most variation in the selection of items from the diverse-culture group to the same-culture group.

The authors assess the advantages and disadvantages of each method and conclude that the TID, chi-square and three parameter IRT are most promising. The major drawback to the TID is that differential performance at different ability levels is obscured with the use of an average p-value. Both of the other methods identify this type of item. Parameterization in the IRT approach and the determination of score levels in the chi-square are problems which need to be solved. The factor score method and chi-square (1-p) approach were

found unsuitable in this study. They recommend doing a similar study again with simulated data so that the item parameters can be controlled.

The purpose of the study undertaken by Ironson and Subkoviak (1979) was to compare bias detection methods which varied in mathematical sophistication, implementation and cost in order to find the most valid method. Also they were interested in the content of the items earmarked as biased. TID, item discrimination, Scheuneman's chi-square and the three parameter ICC techniques were employed to analyze six subtests of the National Longitudinal Study (NLS) of 1972. The sample included 1,691 Blacks and 1,794 Whites randomly selected from 17,726 twelfth graders. Three traditional and two nontraditional subtests were analyzed: vocabulary, reading, mathematics, picture-number letter groups and mosaic comparisons.

The authors computed signed as well as unsigned indices for each method in order to preserve the direction of the bias. A positive sign indicated the item was biased against blacks; a negative sign signified the item was biased against whites. For the TID method, signed and unsigned distance values (d_i) from the major axis were the measure of item bias. For the discrimination index, the absolute value of the

difference between the black and white biserial coefficient was the unsigned bias index and the direction of bias was indicated by attaching the appropriate sign to the differences in discrimination values.

With these two methods, TID and discrimination differences, the signed index preserves not only the direction but also the magnitude of bias. In the chi-square and ICC methods, a signed index is obtained by summing across ability levels. If bias varies across ability levels the discrepancies may be large but compensating so that the signed index is quite small. This study followed Scheuneman's modified chi-square approach which only considers correct responses; signed and unsigned values were included. Finally, the ICC three-parameter procedure used the LOGIST program for parameter estimation and the unsigned bias index was calculated by adding successive rectangles of the area between the curves for the two groups. The signed ICC index followed the same procedure but instead of the absolute value of the difference between curves, a positive sign was attached when the white curve was above the black curve and a negative sign given when the black curve was above the white curve.

Ironson and Subkoviak were interested also in two other comparisons: traditional subtests vs.

nontraditional subtests and a black group and white group who were matched on subtest scores. Because a major criticism of bias research is the confounding that occurs between ability and bias, the authors wanted to investigate bias in two different groups who had the same distributions of ability. The signed white minus black p value difference or the absolute value of the subtest matched p difference were the bias measures for the matched comparison.

The unsigned average bias measure was calculated for each method and each subtest. The most bias was found in the traditional verbal and mathematics subtests by the TID, chi-square and ICC. No pattern was evident with the discrimination method. The largest correlations among signed indices for all the methods in the battery were found for the chi-square and ICC, .485, chi-square and TID, .370, and TID and ICC, .239. All were significant at $p \leq .01$. The unsigned chi-square and ICC indices correlated the highest with the traditional-non-traditional dichotomy. The signed chi-square and ICC indices also showed the biggest correlations with the matched groups index. The authors conclude that the correlation of three of the four bias techniques were significant but small which lends some support to the use of the ICC, chi-square and TID

methods. Again the caveat is included that although the ICC three parameter procedure is recommended, since the cost and data requirements of this method often preclude its use, the chi-square and TID procedures would appear feasible substitutes.

The results of this study are consistent with the findings of the other comparison studies; that is, when bias methods are compared, the ICC three parameter method and the chi-square procedure are most highly correlated, with the TID correlating next highest (Merz & Grossen, 1979; Rudner & Convey, 1978; Rudner, Getson & Knight, 1980a; Shepard, Camilli & Averill, 1980). However, several questions arise from the analyses undertaken in this study. The biggest problem occurs when correlations are given for the battery as a whole for any of the bias measures, but especially for the ICC-3 parameter method. One assumption of the bias techniques is the unidimensionality of the test, that is the test, or subtest, is measuring one underlying ability. This study has combined the Vocabulary and Reading Comprehension subtests into one verbal subtest where it seems different abilities are being measured: the Vocabulary section asks for synonyms of words, while the reading section is comprised of short passages followed by questions. Also, most research which has

used the ICC-3 parameter discusses the difficulties encountered when less than 40 items are included in each subtest (Linn and Harnisch, 1981; Rudner, Getson & Knight, 1980; Shepard, Camilli & Averill, 1980). Only the Mosaic Comparisons have that many items, but it is a speeded test which also poses problems when comparing methods. Another question which arises concerns the matched group analysis. Why did the authors use a heretofore not included bias measure? Since researchers are concerned that differences in the ability distributions of the groups interfere with an accurate bias index, it might have been useful to calculate the eight bias measures (four signed and four unsigned) under investigation in this study to assess the influence of differences in ability between groups on these indices.

More verbal items were identified as biased by all the methods as well as items at the end of the subtests. It may be that the verbal items were selected because the assumption of unidimensionality was not met. The last items may have been identified because the ICC-3 parameter model treats items at the end as not reached and does not include them, but the other methods include these items as incorrect. Douglas (1981) has found that more final items on speeded tests are identified as biased depending on the bias detection methods employed.

Two studies which examined the similarities and differences among item bias detection procedures have included an external criterion as the measure of ability instead of the total score of the test under investigation (Burrill, 1981; Shepard, Camilli & Averill, 1980). The intent of these studies was to avoid the circular reasoning involved when one is looking for biased items in a test while the total score on that test which includes those items is considered the measure of ability.

Investigation of Metropolitan Readiness Test (MRT) item pools was undertaken by Burrill (1981) in her doctoral dissertation. She administered fourteen item pools in three tryout forms from the tryout phase of the 1976 MRT. Three sets of samples were examined: randomly selected blacks and whites, two pseudogroups of whites and a group of whites and blacks matched on an external criterion, which was the total score of the MRT 1965 edition. The external criterion was used as the ability measure only for the matched groups. For the other groups each subtest score was the ability measure. Seven procedures were compared: raw p-values arcsine transformations, delta-difference transforms and the Angoff delta distance; biserial and point biserial correlations; and Scheuneman's modified chi-square. (In

contrast with Ironson's (1979) study, the ICC techniques were not used even though there was a total of 204 items, because the items were not considered unidimensional.) Editorial examination was undertaken of outliers, or items classified as extreme by the bias methods. Rank order correlations were calculated between pairs of procedures for each item pool and for each pair of samples.

Burrill found that all the difficulty procedures were highly correlated with each other, even though Angoff delta distance values correlated less well with the other three. The discrimination indices s correlated very highly across the pairs of samples for all item pools, but almost all correlations of one difficulty index with one discrimination index were nonsignificant. The correlation between a discrimination index and the modified chi-square were only modest; while the correlation between a difficulty index and the modified chi-square ranged from .62 to .78. One important finding was that matching groups on an external criterion improved the correlation in all cases but to a different degree.

Items selected by each method for each sample were compared to see if the same items were identified in the random samples of black and whites and the matched

black and white sample, but not in the two random white samples. Extreme items in each method and for each sample were defined as those items more than 1.5 standard deviations from the mean of the distribution of the signed item index. Also, the items which ranked first, second, next-to-last and last in the distributions were examined. No procedure was found to consistently identify the same items as biased for both the matched and random different ethnic groups, but not select those items for the random white samples. Twenty-four items were classified as biased by this scheme for all the procedures.

No conclusions could be reached about the content of the items identified as biased by any procedure, although more verbal items were classified than those requiring auditory discrimination. The author states that format such as position on a page or the position of the correct answer "may be much more of a problem in creating artificial bias indices" (p. 141). Also, a preponderance of items were classified as biased which contained negatives in the question or response alternatives.

The problems were discussed which occur when signed indices are compared across methods when the modified chi-square procedure is included. As mentioned

previously, signing preserves the direction of bias and improved almost all the correlations among methods in the Ironson & Subkoviak study. However, in the chi-square method, differences in response patterns for different groups are minimized when these differences are summed across ability groups. Burrill discovered that the rank order of the signed chi-square values was often different from the rank order of the other methods by one. Although in one case, a small difference changed the ranking of the chi-square value from four to 18.

The Raven's Coloured Progressive Matrices was the external criterion used by Shepard, Camilli & Averill (1980) in their study which analyzed the Lorge-Thorndike Intelligence Test, Verbal & Nonverbal, Level 3, Form B, 1954 for item bias. This research had several purposes. First, the authors wanted to determine if the results from the ICC 3-parameter (ICC-3) method could be approximated by any other simpler method. Secondly, by including an external criterion, the validity of the various methods could be examined. Last, the items selected as biased when the total score of the Lorge Thorndike verbal or nonverbal was entered as the ability measure, were to be compared with those items chosen when the Raven's total score was substituted.

Six item bias detection methods were considered with a total of 16 bias indices calculated: 1) TID-Angoff computed the perpendicular distance from the major axis of the point on the bivariate graph which represents the paired transformed difficulty values for two groups; 2) point-biserial difference between two groups; 3) ICC-3 used the LOGIST program to calculate the difference in discrimination parameters (a's), difference in difficulty parameters (b's), the unsigned and signed area between the curves for two groups and the composite test of differences in both a and b parameters; 4) ICC-1 method obtained unsigned and signed area measures, differences in difficulty (b) parameters and a weighted difference in difficulty parameters which takes the variance of the b's into account; 5) Scheuneman's modified chi-square found unsigned and signed values; and 6) Camilli's chi-square specified unsigned and signed values.

Three almost equal groups of randomly selected fourth, fifth and sixth graders were selected from data originally analyzed at the total score level by Dr. Arthur R. Jensen. Included were 490 black, 551 Chicano and 552 white pupils from lower to middle class homes. The two tests were administered the same week to these examinees in their classrooms as part of a battery of

tests given during that week. Comparisons were made for the black-white and Chicano-white samples for each analysis. All methods were included when the Raven's score was substituted for the total score on the Lorge-Thorndike except for the TID which does not lend itself to such computation. The Raven's was chosen as the external criterion because it was considered less culture-loaded than traditional intelligence tests such as the Lorge-Thorndike. Since the Raven's is more similar to the Lorge-Thorndike nonverbal test, the authors believed that results would be more similar when the Raven's was substituted for the nonverbal section than the verbal section. This hypothesis was confirmed; especially for the Chicano-white analyses.

As expected, the correlations between different indices of the same method often reflected close agreement. However, the ICC-3 b differences measure did not correlate with the a differences measure; neither did the ICC-1 fit statistic correlate with the ICC-1 b difference index. On the other hand, correlations between signed and unsigned values of the same method were low because signed extreme values are at two ends of the distribution and the unsigned values represent only one end. Almost perfect correlations (.99) were found between the TID Angoff technique and the ICC-1

method measured by the difference in b values and signed area. The authors state that since these two methods appear so similar, the TID Angoff procedure should be preferred because it is computationally much easier and can be explained graphically. The ICC-1 fit statistic did not correlate with the other ICC-1 indices and may be measuring something quite different. Point-biserial differences, ICC-1 fit index and ICC-3 a differences stand by themselves; no relationship seemed to exist between these indices and the others.

Biased items were defined as an "anomaly in a context of other items" (p.4); that is, the item by itself is not biased, but rather a particular item within a group of items is not consistent with the set of items. Shepard, et al. believed that items classified as extreme in other studies were often on one side of a cut-off point, while other items not so classified were adjacent on the index. In order to be labelled extreme in their study, items had to be separated from the main cluster of items by gaps in the distribution. Outliers were chosen for each index by inspection of histograms of the distribution of item bias measures. Those items selected were removed from the homogeneous and uninterrupted majority of items for each index and therefore, the same number of items was not selected for each index.

Although this makes comparisons among indices more complicated, results are more meaningful. Again, more agreement was found among different indices of the same method than between different methods, but five of the ninety items were selected by three or more methods. A table of all items on the verbal Lorge-Thorndike which shows which items were identified as extreme by which indices clearly portrays from the Black-white comparison how the vast majority of items were not classified extreme by any index.

The original intent of this research was to compare other bias detection methods to the ICC-3 techniques, which was considered the method of choice. According to the authors several problems arose in the computation of this measure which do not make these results representative of the ICC-3 method at its best. First, the sample sizes did not approach the 1,000 subjects recommended for analysis. Not only was convergence of parameter estimation difficult for _c_ values, even though averages were used, but also for the _a_ and _b_ parameters. This was especially true when the external criterion was employed. The equating technique presented difficulties too. The ICC's were obtained separately for each group, then a principal axis line of best fit was derived first for the _b_'s and then for

the a's. Shepard, et al. state that if the parameters are estimated well, this procedure might be satisfactory, but in their case with poor estimation a few outlying items made the principal axis unstable. Finally, as stated previously, the LOGIST program for the ICC-3 method treats omitted responses as different from wrong answers, while the other methods treat omitted answers as wrong. This research counted omitted answers as wrong also for the ICC-3 technique to control for differences in methods. Examinees were eliminated however, if they did not respond to at least one-third of the questions. The authors were unable to determine how this change in LOGIST affected the outcome and plan in the future to rerun the program with omitted responses considered as omitted responses.

A ceiling effect was demonstrated on the Raven's for the white sample which weakened the full impact of this test as the external criterion. The correlations of each bias measure with itself using the internal then external criterion were almost perfect for all the ICC-1 indices except the fit statistic. Since ability differences were smaller on the external criterion, models correctly identified more items as biased. The ICC-3 procedure was sensitive to the differential relationship between the two Lorge-Thorndike

subtests and the Raven's. Much less agreement was apparent between the Raven's and the Lorge-Thorndike verbal subtest.

A major conclusion of this study is that the method one chooses makes a difference in the number of items selected as biased. Enough convergent validity was demonstrated, however, to "hearten a measurement theorist" (p. 74), with the relationships among indices indicating they are tapping the same psychometric property.

Several studies have used simulated data to introduce varying amounts of bias a priori and then have compared the efficacy of various bias detection methods in identifying induced bias (Merz & Grossen, 1979; Rudner, Getson & Knight, 1980a). Obviously only item statistics can be compared, since there are no actual items to inspect, but hypotheses can be tested about the effect of differences in item discrimination or difficulty for different groups on the bias techniques currently in use.

A Monte Carlo procedure specified a priori the amount and type of item bias in the Rudner, Getson and Knight study (1980a). Test length, amount of bias in discrimination and amount of bias in difficulty were

manipulated to produce 112 different combinations of test conditions. Two item difficulty approaches were used: Angoff's TID which calculates the perpendicular distance from the major axis line and a procedure which directly compares items difficulties. This is done by computing item p-values for each group and transforming them to within group z-scores using the mean and standard deviation for that group. For this index the 45° line is considered the point of item deviation. The single parameter ICC (ICC-1) method was calculated with two indices, the fit statistic and absolute differences in item easiness. For the ICC-3 procedure, Urry's (1975) iterative minimum chi-square technique estimated the parameters and the bias measure was the absolute difference of the area between the curves. Scheuneman's modified chi-square method was followed with values calculated for the traditional five score intervals as well as for as many score intervals as possible total score values, minus the number of cells with expected values less than five.

Since the ICC-3 method both simulated the test data and detected item bias the highest correlations between generated and identified bias were found for this technique. Interestingly, the ICC-1 fit statistic, which is only concerned with the difficulty of an item,

appeared more sensitive to the differences in the discrimination of an item. Test length did not seem to be influential for any index. When differences in item discrimination increased, the correlations between generated and detected bias remained the same for ICC-3; decreased for the chi-square, TID indices and ICC-1 easiness index while increasing somewhat for the ICC-1 fit index. For changes in the b parameters, or item difficulties, correlations between generated and detected bias for the ICC-3, ICC-1 easiness index and TID-45° index display a steady rise and the ICC-fit measure shows a decline.

The authors state that although the ICC-3 method was used to both simulate and detect bias, the high correlations found between generated and identified bias demonstrate the greater accuracy of this method over other methods. The correlation for the ICC-3 was .80, however, which does not seem that high when this was the method which introduced the bias. Higher correlations were not obtained for this method because of difficulties in obtaining parameter estimates for all items and consequently some items were excluded from the analysis. The close correspondence of the correlations of Scheuneman's modified chi-square procedure with five

score intervals to the ICC-3 correlations further substantiates the similarity of these methods.

Merz and Groseen (1979) wanted to investigate the statistical procedures used to detect item bias by systematically varying the difficulty parameter of simulated data of two hypothetical groups of examinees. A Monte Carlo procedure was used to simulate data according to Birnbaum's ICC-3 model and specified hypothetical data with known parameters for two groups of 1,000 examinees each. Two levels of difficulty on a sixty item test were generated, 60% and 80%, with no biased items, 10% biased items and 20% biased items in each condition. Discrimination parameters and guessing were held constant across conditions at 1.0 and .05 respectively. The definition of bias was an area of .70 or greater between the ICC's for the two group.

Six bias detection procedures were followed with six bias measures computed and then converted to z-scores for comparison. The ICC-3 and ICC-1 methods calculated the area between the curves; for the point-biserial index the difference between item-total correlations for the two groups was used; and the TID method designated the distance of the point of paired z-values from the 45° line as the measure of bias. The chi-square procedure set eight-score intervals for the

60% difficulty conditions and six score intervals for the 80% difficulty values, with the bias index the signed chi-square. No rationale is stated for the number of score intervals used in this study which differs from other research. Rudner, et al. (1980) found that increasing the number of score intervals reduced the accuracy of the chi-square method, when compared to other methods or the chi-square method with five score intervals. Factor analysis entered group membership as dummy variables into the item intercorrelation matrix, then that matrix was reduced with Principal Components Analysis and rotated with varimax rotation (Merz, 1976). Items which had the largest factor loading on the same variable as group membership were considered biased. No value was specified by the investigators.

A major problem confronted by researchers in item bias is the confounding that occurs when the differences between two groups in performance are an indication of the differences in ability rather than bias. This study defines bias as the differential probability of a correct response for persons of equal ability. While intuitively appealing, the problem remains how to separate the ability measure from test performance. In this case, the average per cent correct for the total score was kept equal by biasing 50% of the

items in favor of one hypothetical group and 50% in favor of the other hypothetical group. Thus, total score distributions and test characteristic curves were kept similar. While this similarity is unlikely in actual testing situations, the confusion is reduced between differences in ability and bias so that the comparisons of the bias detection methods can be explored fully.

Correlations between generated and detected bias were determined for each method. The TID approach had the highest correlations and was uniformly consistent in identifying biased items under all conditions, while point-biserial correlations were uniformly low. The correlations for the other methods were high but demonstrated erratic fluctuations at certain conditions. For instance, the factor score correlations were all above .92 except for the 60% difficulty, 10% bias condition where it was .10. No apparent reason could be found for these discrepant results.

The accuracy of each method in classifying the appropriate number of items as biased was assessed. Most of the methods over-identified items are biased: ICC-3, ICC-1, TID and factor analysis. Point-biserial underidentified items and the chi-square overidentified in two cases, while it underidentified items in two other

cases. The various methods selected items correctly in varying degrees. Again, TID was the most successful.

The results of this research differ somewhat from the findings of other similar studies in that the TID method appeared the most accurate in detecting induced bias. Logically, this method, which only examines the differences in difficulty values for the groups, would seem most sensitive to manipulation of the difficulty parameter. However, the ICC-1 method also is a measure of the differences in the difficulty values of the item and it performed less well. Again, greater concurrence then was detected by the ICC-3 procedure was expected since the bias was introduced with this method. Also, no mention is made whether the ICC-3 or chi-square procedures portrayed items with differential validity. Even though the data was generated so that ability and bias would not be confounded, when the methods were used, did the "performance" on an item change at different ability levels? This would provide evidence for a method effect which may be part of the influence of the differences in the abilities of the groups.

The comparison studies which examined actual test data had no a priori hypothesis about the number of items which would be classified as biased. Consequently, the validity of the techniques could only

be compared to one another, not to any value of true bias. On the other hand, the results of the simulation studies which did determine the extent of bias a priori are confounded because the method used to simulate bias in an item was also used to classify items as biased by the same parameters. In order to eliminate this confusion and to use real data Subkoviak, Mack and Ironson (1981) conducted a study whereby items which asked for definitions of black slang were included in a standardized test of vocabulary. The degree of correspondence with the a priori bias was calculated for ICC-3, Angoff TID approach, and chi-square index, both Scheuneman, which utilizes only correct responses and Camilli's which considers correct and incorrect answers. The ICC-3 was found to be the most valid procedure in detecting a priori bias, with a correlation of .87. The authors contend that the other three methods which had high intercorrelations of .90, .94 and .97 appear more similar to each other than the ICC-3 and thus may be measuring something different.

This study does not discuss the items selected as biased by each method nor do they consider the possibility of bias within the standardized test. Also the correlations are computed over the entire set of items. It might be of interest to test developer and

reseacher alike to know how many items, both new and old, correlated perfectly with the different methods and how many items were not classified as biased at all.

A pervasive thread which runs through the current research on bias procedures is the difficulty encountered by actually implementing the ICC-3 method. Investigators would like to use the most sophisticated methods available, such as the ICC-3 procedure, and many feel that the problems which arise are only nuisances to be eliminated through educated use. However, aside from the oft mentioned data requirements and computer time necessary to obtain parameter estimates, several major difficulties remain if this technique is to provide meaningful research data on item bias. First, the loss of items is not addressed in any study. While the numbers vary for each study and for each test in each study, generally about ten per cent of the items are excluded from the analysis. Usually the reason given is wildly fluctuating parameter values, but the effect of the loss of items has not been investigated. Studies which compare methods would seem especially influenced by the exlusion of items for one method and a review of these items might provide useful information.

Second, what should be the measure of bias when the ICC-3 is used: the area between the curves, the

differences in the discrimination parameter (a's), the differences in difficulty parameter (b's), or the significance test applied to the variance-covariance matrix of the a's and b's? In the only study to use more than one ICC-3 index, Shepard, et al. (1980) found that the significance test index correlates highly with the chi-square methods (Scheuneman & Camilli) and the TID-Angoff. The ICC-3 b difference value and area indices correlate with one another but it could not be determined if they are classifying the same items or not. Other studies, which have compared the area index with other methods, have found much concurrence between the ICC-3 area and the chi-square technique (Ironson & Subkoviak, 1979; Merz & Grossen, 1979; Rudner, Getson & Knight, 1980). Wightman (1979) stated that the area measure was caluclated across ability levels which is somewhat misleading. The area of interest should be in the center of the distribution for each group.

Equating procedures have also proved troublesome. Linn and Harnisch (1981) suggest deriving ability estiamtes with the total sample, then dividing the ability scales into quintiles. Within each quintile comparisons between the observed score with the estimated score for each group are found and a difference index is calculated.

Lord (1980) outlines a complex strategy for equating parameters of different groups which has not been tried by any bias study to date. It is analogous to Sinnott's (1981) procedure to derive the major axis by removing items with extreme bias first. In Lord's approach, the whole test is analyzed first for each group and items with significant response functions are removed. Then the groups are combined and ability estimates for the whole group calculated. The a and b parameters are now estimated separately for each group and each item, including the items which has been removed. Finally, a chi-square test which examines the variance-covariance matrix of the a and b's is computed for significance.

The research on item bias techniques appears to attest to the similarity in the results found with the ICC-3 and chi-square methods. The chi-square procedure which uses correct and incorrect responses can evaluate all items on all tests with relatively small sample sizes. The chi-squares can be compared directly if the same score intervals have been set for each item. Also patterns of differential validity within the item can be assessed. It remains to be investigated whether including the total score as the ability measure in the chi-square method is more serious than the problems encountered by the ICC-3 with parameter estimation.

Two recent simulation studies have varied one or several item parameters as well as the amount of bias in a test in order to examine the interaction of these parameters on bias measures (Ironson & Craig, 1983; Scheuneman, 1982). Scheuneman used the three-parameter IRT method to alter item difficulty, item discrimination, item distributions and amounts of bias. The hypothetical groups had an ability difference of one, with identical total score distributions. In this simulation it was found that a peaked distribution resulted in larger true score differences than a uniform distribution. Also, tests with low discrimination had larger score differences; while tests with medium or high discrimination demonstrated fewer inflated differences because of bias. This study focused on item parameters which appeared most likely to influence item bias and consequently which parameters bias methods need to identify. For instance, item difficulty, the parameter to which most detection methods are sensitive, appeared most influential in score change. However, Scheuneman cautions against the use of techniques which do not consider item discrimination or guessing, which were found to affect the classification of an item as biased also.

In another simulation, Ironson and Craig (1983) altered the amount of bias on a test as well as ability differences between the two groups to investigate the influence on test reliability and validity and to access the agreement among four bias methods (TID, chi-square, three-parameter and one-parameter IRT). High agreement was found among the bias methods for all conditions except large bias and large ability differences. Under all conditions reliability remains stable while validity decreased somewhat as the number of biased items increased. The mean score differences between the two groups increased as the amount of bias increased. These two studies both highlight the fact that as more information about the influence of item parameters on item bias is known, the more information can be used at the test development stage.

Studies which have investigated methods of detecting bias in the individual items on a test have attempted to analyze the items which have been classified as biased to determine the reason that a particular item or set of items has been chosen (Burrill, 1981; Ironson & Subkoviak, 1979; Linn et al., 1980; Linn & Harnisch, 1981; Nungester, 1977; Rudner & Convey, 1978; Scheuneman, 1975, 1976, 1979, 1980, 1982; Shepard, Camilli & Averill, 1980; Stricker, 1981). The focus of these analyses has

been on item content and item type. Specifically, did the items selected as biased differ in content and type from the other items not identified as biased? Stricker conducted the only study which classified items first, and then after biased item were identified, computed correlations between the items chosen as biased and the item type, i.e., analogies or quantitative comparisons. He found no correlations between item type and bias but did find that verbal items with female content favored females. He concluded that the classification scheme was not sensitive enough to the different types of items.

More verbal items have been found biased against blacks and males, than nonverbal items (Blew & Ishizuka, 1978; Blew & Stern, 1979; Breland, 1974; Burrill, 1981; Draba, 1977; Rudner & Convey, 1978; Scheuneman, 1979). However, on an experimental social studies test as part of ACT assessment tryout, verbal items seemed to favor males and whites (Huntley & Plate, 1980). Of the verbal items classified as biased, analogies have consistently been found to represent a greater percentage of items selected than any other item type. However, when analogy items designated as biased are compared with analogy items not selected as biased, no differences seem apparent (Stricker, 1981).

The effect of content relevant to one group, such as males or females, has been studied with conflicting results. Medley & Quirk (1974) altered the content of the National Teacher Examination and found blacks and women improved their scores when the content was relevant to those groups. McCarthy (1976) discovered that the item statistics of a math test varied for males and females, depending upon whether the item content was expressed in male, female or neutral terms. Real world referrants in an item appear to favor males in one study (Strassberg-Rosenberg & Donlon, 1975), while demonstrating no effect in another (Stricker, 1981). Schmeiser (1980) conducted an experimental study of several forms of an English and a social studies test each of which contained content relevant to either black culture or white culture. Everyone took both English tests, but only one social studies test. No differential effect was found for test content on examinee performance by race. In a follow-up study to control for skill level, Schmeiser devised two tests with black content and two tests with white content, then administered them randomly to an equal number of blacks and whites. Whites achieved higher scores regardless of the content of the test. The author recommends that test content be as diverse as possible to reduce the influence of content relevant to only one group.

A different conclusion based upon the same information was advanced by other studies. Intelligence test content was analyzed by Zoref and Williams (1980) for evidence of ethnic and sexual sterotypes. They reviewed the most widely-used tests including the "culture-fair" ones and concluded that there is a serious imbalance in these tests towards white male superiority. The presence of subtle forms of biased content, such as sterotypes should be excluded from tests, even if no performance differences exist. The question which arises from these studies is whether the identification of races or sexes introduces a nuisance factor or does it motivate the particular group mentioned, especially when the reference is irrelevant to the skill one wishes to measure (Haebara, 1980)? Huntley & Plake (1980) also found whites did better on their experimental test forms which included many items which specified black or black content. Instead of balancing the content to include all groups they suggest items be phrased as neutrally as possible though this is difficult to achieve.

Lloyd (1982) analyzed the ACT assessment test to investigate item bias against Hispanics. She used a three factor ANOVA design (items, sex and ethnic group) on four subtests. While significant interactions were found for several items the bias was balanced between

Whites and Hispanics. The author felt that there was little, if any, difference in the total scores of each group, and therefore the biased items were inconsequential.

Item properties which introduce the element of uncertainty for the examinee may produce biased items. Instructions for item writers and reviewers of items stress the importance of clarity. It is germane to the purpose of testing that an item be written and presented with no ambiguity for the examinee who possesses the ability being measured. In other words, all examinees who attempt the item should have the same chance of success or failure based upon their ability, not upon their sophistication to answer certain types of questions, or awareness of clues which would lead to the correct response. Perfect items are an ideal that test companies would like to believe can be developed through rigorous review, but it is likely that some flaws remain to which groups differ in their sensitivity (Roid & Wendler, 1983; Scheuneman, 1982).

Directions on achievement tests are usually read aloud to examinees; then, a practice item is completed together to illustrate what is expected and questions are answered before the testing begins. Often

only one sample question is presented for an entire subtest, e.g., reading comprehension, but occassionally more than one item is given at the start and another practice item included if the format changes. For an examinee with limited testing experience, the sample, or practice, item is the only exposure to the demands of the test. When only one type of item is used in the sample, with other item or response formats used in the test, the first item of a new format may be approached with trial-and-error. In fact, the item may be considered a practice item because the examinee first needs to understand the requirement of the question before a suitable answer can be found (Scheuneman, 1976).

Studies have demonstrated the need for clearer directions and more practice items (Huntley & Plake, 1980; Oosterhof, Atash & Lassiter, 1982; Scheuneman, 1982a). During the item tryout phase of the Otis-Lennon Metal Ability Test for junior and senior high school, many analogy items were classified as biased against blacks. When a sample item of that format was included in the test, most of those items previously identified were no longer identified. Huntley and Plake (1980) hypothesized that the lack of sufficient practice may have been the reason several first items in a section were flagged as biased.

Using the Angoff transformed item difficulty techniques to produce a bias index for each item, Oosterhof, Atash and Lassiter (1982) graphed groups of items in order to test hypotheses about test directions and time limits. They felt that directions which were unclear for one group could be expected to result in fewer beginning items answered correctly for that group. The Flight Aptitude Selection Test was analyzed with White males constituting the majority group and Black males, Hispanic males and females comprising the three minority groups. When the first few items were graphed, it could be seen that the minority groups did perform less well on the start of a section. This suggested to the authors that clearer directions or more practice items may aid minority examinees to understand the demands of the question before the test begins.

Questions which are written in the negative have been selected as biased more often than items written in a more straight forward manner (Burrill, 1981; Scheuneman, 1978). These items are expressed contrary to usual questioning (e.g., One reason Sam did not open the door?), and may perhaps require a different type of judgment or awareness. Uncertainty and ambiguity can arise with wording such as IF/THEN, EXCEPT, EITHER/OR, LEAST, PROBABLY, POSSIBLY or DIFFERENT FROM.

Response alternatives which include ALL OF THE ABOVE, NONE OF THE ABOVE, or ask for some combination of answers not only have higher difficulty values (Dudycha & Carpenter, 1973; Hughes & Thumber, 1965; Tollefson & Tripp, 1982) but also appear in many items selected as biased (Huntley & Plake, 1980). According to one test publisher's suggestions to item writers, this type of response alternative should be avoided and questions which require one best answer seem preferred.

Tollefson & Tripp (1983) conducted two studies to investigate the effect of including NONE OF THE ABOVE as a response option. In the experimental study this option was systematically included as the correct answer, included as a foil or not included. Items which had NONE OF THE ABOVE as the correct answer were found to have a significantly higher mean discrimination and were somewhat more difficult.

Ambiguity can also be introduced into the question with an inadequate item stem (Crowder, 1979; Scheuneman, 1976). If the question is a short incomplete stem, the information may be inadequate for the examinee to choose an answer. When Scheuneman (1976) elaborated an incomplete item stem which had been identified as biased on the tryout form of the Metropolitan Readiness

Test, it was not classified as biased on the final form. Both the lack of sufficient information when too few words are presented and the complexity of the language contained in an incomplete sentence may hinder an examinee from fully understanding the question (Crowder, 1979). An incomplete stem as a question may not be as clear a presentation while a question with two or more sentences may be confusing.

On a mathematics subtest, word problems demand that the examinee decide first on the mathematical process required before computation can be attempted. Therefore, computation items would seem more clearcut for the examinee than word problems where vocabulary, reading ability and abstract reasoning are needed before computation. Also, on a science test, a skill additional to the knowledge of science would seem necessary in order to understand visual content such as graphs, charts or diagrams before the knowledge of science can be used.

In reading comprehension, it is possible that the length of the passage may interact with group performance. Several studies have found that black children omit items instead of quessing or will select the first logical response alternative rather than read through the list of alternatives (Burrill, 1981; Frary & Giles, 1976; Scheuneman, 1979). This implies that

first alternatives may be selected more frequently than the last alternative, especially on difficult items. Also on comprehension subtests with questions which refer to passages, the longer the passage the less likely some examinees will read the passage through in order to find the part related to the question.

Differences in the length of the response alternative may be inadvertent clues to examinees (Ebel, 1972). In some cases short alternatives may be quickly eliminated as not containing enough information, while on other items the shortest answer may be the only clearly stated alternative. When all the alternatives are the same length, this confusion may be reduced.

Finally, the actual position of the item in the subtest appears to affect item parameters, such as difficulty values, as well as differences in group response patterns (Flaugher, Melton & Myers, 1968; Whitely & Dawis, 1976; Yen, 1980). Whitely and Dawis (1976) designed seven tests with the same core of 15 items and 45 different items which were administered to groups of approximately 200 examinees. They wanted to test for differences in Rasch and classical parameters in the fifteen core items located differently in each test. Nine items had statistically significant differences in classical difficulties and 6 of the 15

items had statistically different Rasch difficulties. A core of items in different booklets was used by Yen (1980) to explore the causes of contextual differences in item parameters using both the one and three parameter methods. It was found that items placed at the beginning or the end of the booklet showed the most variability. The items at the end appeared more difficult than when the same items were at the beginning of the test. The author conjectured that the examinees were not as careful at the end of the test as at the beginning. While the discrimination of the item varied more than the difficulty values, the changes were not systematic.

Research has shown that statistical measures of item bias are influenced by observable characteristics of items. Based on the current research the present study tried to define and quantify a set of item properties which would predict a measure of item bias and which could be used by test developers to assign an index of bias to an item.

CHAPTER III: STATEMENT OF PROBLEM

This study defined item properties which previous research suggested may be predictive of item bias. These item characteristics were then scaled on the Reading, Math and Science subtests of the Metropolitan Achievement Test, 1978, Form J (MAT-J), and correlated with various statistical indices of item bias.

On each subtest a set of seven item properties were examined. For the Reading subtest these included: 1) Reference to any group 2) Similarity to practice item 3) Length of item stem 4) Difference in the length of the response options 5) Question format 6) Item location and 7) Passage length. The item characteristics which were defined for each item on the Math subtest were: 1) Similarity to practice item 2) Length of item stem 3) Difference in the length of the response options 4) Question format 5) Response format 6) Arrangement of response options and 7) Type of problem. On the Science subtest each item was scaled on the following properties: 1) Length of item stem 2) Difference in the length of the response options 3) Arrangement of response options 4) Total number of words in response alternatives 5) Number of words in correct answer 6) Item location and 7) Presence of visual material.

The full chi-square index was used as the measure of item bias and developed for each item as both a continuous and dichotomous value. In addition, four different methods for constructing the majority (i.e. White) and minority (i.e. Nonwhite) groups were employed. By considering the two different chi-square measures (continuous, dichotomous) for each of four grouping procedures, eight different statistical measures of item bias were developed for each subtest.

As previously noted the classification of items as biased depends upon the method which is used. The full chi-square index was selected as the measure of item bias for several reasons. First, in comparative studies of item bias methods, the full chi-square method correlates most highly with the three parameter IRT method. Whereas the IRT method is considered the most sophisticated and technically advanced, the sample size and number of items required for this method were not available in this study for all samples and subtests. Second, all the items on each subtest were to be examined which is not always possible with the three parameter IRT. Third, the chi-square method does not make assumptions about the shape of the distribution of scores which can be a problem with the transformed item difficulty methods and discrimination methods. Fourth,

there is some control for differences in the abilities between the groups by setting total score intervals. While the chi-square method possesses these advantages several disadvantages must be considered. The major disadvantage is the influence of sample size on the chi-square, which increases as the sample size increases. This is of greatest concern for the dichotomous chi-square(DCS) which has been defined as a continuous chi-square(CS) that is significant at $p<.05$. Also, the chi-square method is not sample invariant, but will be influenced by the sample and the score intervals.

Multiple regression was performed to assess the degree to which the item properties predicted the eight chi-square measures. The sets of regression weights derived from the MAT-J were cross-validated on the Metropolitan Achievement Test, 1978, Form K (MAT-K), and the General Educational Development Test (GED). The use of the MAT-K enables one to consider the generalizability of the prediction equations to another test form with a similar age sample of examinees, whereas the use of the GED extends the generalizability to a different test and a different age sample of examinees.

The following questions were explored:

1) Will some combination of observable item properties predict item bias?

2) Will the same or a different combination of item properties predict item bias in the different content areas: Reading, Math and Science?

3) Can prediction equations be developed which can be generalized to other tests and other samples to identify the degree to which item bias is present?

CHAPTER IV: METHODOLOGY

This study proceeded in three stages. Part I first analyzed data from the standardization sample of the Metropolitan Achievement Test, 1978, Form J (MAT-J), for grades seven, eight and nine to establish the criteria by which the item characteristics would be scaled. Next, eight different measures of item bias were determined for each item. These measures were constructed by considering both a continuous and dichotomous form of the full chi-square measure for each of four different procedures for grouping the majority (i.e. White) and minority (i.e. Others) examinees. Then, prediction equations were derived based upon the multiple correlation of these characteristics with the eight chi-square indices. Part II used these prediction equations for cross-validation on Form K of the MAT, 1978 (MAT-K), with a different sample of examinees of the same age. Part III is an additional cross-validation that analyzed test results from the General Educational Development Test (GED), which is a high school equivalency test given to examinees ages 17 and older.

DESCRIPTION OF TESTS (MAT-J, MAT-K, GED)

The Metropolitan Achievement Test, 1978, Forms J and K for the seventh, eighth and ninth grade contains

five tests: reading, mathematics, language, science and social studies. In this study only the reading, mathematics and science subtests were analyzed as representative of a diverse range of test content. The 1978 MAT underwent major revisions from previous versions in order to make it more representative of the junior high curriculum throughout the country. Content specifications were outlined and items were designed to assess definite objectives. The test is not considered a speeded one because the time limits are generous; however, a small percentage of examinees did not complete every subtest.

The reading subtest comprises 55 items based upon passages of various length and reading level with six, or in one case, seven different types of questions about the passage. Some questions ask the examinee to select the best title for the passage, some ask for information stated exactly as it is written in the passage, while other questions want the examinee to infer from the passage. Also included are questions which ask for the definition of an underlined word as it is used in the passage. An effort has been made to use vocabulary familiar to junior high students who could read at that level, but the topics of some passages appear quite removed from the experiences of most junior high

students. For instance, the first passage discusses how marmalade got its name from the French Marie est malade, "Mary is sick."

The mathematics subtest contains two parts but within one time limit. The first section consists of word problems, number applications and numerical concepts for which there are four numerical response alternatives. The second section has 17 computation problems with three numerical response alternatives and NG, or not given, as the fourth alternative.

The science subtest covers earth science, physics, chemistry, biology and ecology. Many questions are based upon some visual material such as graphs, charts or diagrams. Questions which require inferences as well as rote information are presented. One graph, which portrays the weight of twins as they grow from 6 to 18 years, has questions which appear to assess whether the examinee can read a graph and is very similar to questions about a graph in the math subtest.

Form J, and Form K, are considered parallel forms of the 1978 Metropolitan Achievement Test. In Special Report Number 11 the reliability estimates and standard errors of measurement are presented. Kuder-Richardson Formula 20 was used to compute reliability

estimates which are similar for the two forms. For grades 7 and 8 the coefficients for the two forms for Reading range from .93 to .94; for Math, the range for Form J is .87 to .89 and Form K .88 to .92; and for Science the range is from .86 to .88. The standard errors of measurement in terms of raw score units are also similar for the two forms. On the Reading subtest Form J has a range of 2.7 to 3.1 raw score units and the range for Form K is 2.9 to 3.1. For the Math subtest Form J is 3.0 raw score units for grades 7 and 8 and for Form K the range if 2.9 to 3.1. The standard errors of measurement for Science are the same for both forms across grades 7 and 8.

The GED is comprised of five tests: Writing Skills, Social Studies, Science, Reading Skills and Mathematics. The information booklet sent to prospective candidates comments that since most examinees are adults who have been out of school for some time the test tries to assess general knowledge rather than specific facts in these areas. Again, only the Reading, Math and Science subtests were analyzed. The Reading test contains 40 questions based upon passages, poetry and an advertisement. The questions appear to require comprehension, vocabulary and inferential reasoning. The Math test asks fifty questions which cover computation,

word problems, graphs and charts. There are sixty items on the Science subtest which consists of a wide range of scientific content such as biology and physics. Both factual information and scientific applications are required. Many questions refer to visual material. Each question has four response alternatives.

The MAT-J, MAT-K and GED are all achievement tests designed to measure skills in several subjects although only the Reading, Math and Science subtests are examined in this study. The MAT-J and MAT-K are administered to the same narrow range of the population, namely students in grades 7, 8 and 9, whereas the GED is given to examinees ages 17 to 76. The MAT-K was selected because it is a parallel form of the MAT-J, taken by a different sample of the same age students. The GED was chosen for cross-validation because while it is similar to the MAT-J in content, although the range of material covered in the subject areas is much broader, the age of the examinees is different.

DESCRIPTION OF EXAMINEES

The examinees who took the Metropolitan Achievement Tests, Form J and K participated in the standardization study of the 1978 revision of that test.

Although about thirty thousand seventh, eighth and ninth graders were administered the tests, only approximately eighteen thousand examinees are included in this study. First, the ninth graders were not considered because a ceiling effect may have been present. Second, it was required that the students complete at least half the questions in each subtest. Third, biographical information such as ethnicity and sex was needed in order to define the groups. About half the examinees were male and half were female, while 25% of the seventh grade and 15% of the eighth grade were other than White.

Table 1 presents the ethnic composition of each test. Because there were not a sufficient number of examinees in each minority group to analyze, the groups were combined into an Other category. The total number of examinees for the MAT-J used in this study was 9375, and for the MAT-K 8930 examinees were used in the total sample. The total number of subjects included in this study varied for each GED subtest: Reading, 1382 examinees; Math, 1355 examinees and 1376 examinees took the Science subtest.

Table 1

Ethnic Composition of Examinees

	TESTS				
	MAT-J	MAT-K	GED		
			Reading	Math	Science
White	7620	7056	1047	1025	1046
Others					
Black	1235	1351	190	197	194
Hispanic	308	373	69	64	66
Orientals	20	32	*	*	*
Indians	82	60	*	*	*
Others	86	58	76	69	70

* **Note:** Included in Others

CONSTRUCTION OF THE EIGHT CHI-SQUARE INDICES

The Continuous Chi-Square (CS) and Dichotomous Chi-Square(DCS) Measures

Item analysis were performed on each subtest, Reading, Math and Science, as a separate unit: the total subscale score, or number right, was the total score on that subtest. Items not answered by an examinee were counted as missing and dropped from the analyses. Often these items are considered wrong in item analysis, but for the purposes of examining item bias only items actually completed were included.

The full chi-square procedure described earlier was used to calculate a measure of bias for each item (Alderman & Holland, 1980; Shepard, Camilli & Averill, 1980). An unbiased item is defined as one where the probability of a correct response is the same for all examinees of a given ability irrespective of group membership. The measure of ability for this method is the total subtest score, or total test score. It is assumed that the test is valid, reliable and unidimensional. No assumptions are made about the shape of the distribution of subtest score or total score.

In order to control for ability level, the observed and expected frequencies of item responses across groups are evaluated at discrete total score intervals. Generally, the literature supports the use of three to five intervals depending on the sample size (Rudner, Getson & Knight, 1980; Scheuneman, 1982; Shepard, Camilli & Averill, 1980). In a comparison study of statistical procedures to detect item bias, which considered both five and eight score intervals for the chi-square method, it was found that the chi-square method with five intervals was most similar to the ICC-3 (Rudner, Getson & Knight, 1980). Alderman and Holland (1980) found interpretable results in analyzing the Test of English as a Foreign Language with ten score intervals but no comparison can be made since different intervals were not used. Except where otherwise noted, five score intervals were maintained in this study.

To establish score intervals the total score distribution of each subtest is inspected; then, the sample is divided roughly so that there will be an equal number of examinees at each score level. That is, the White group is divided by five and the Other group is divided by five so that each group is divided somewhat evently. This does not mean that within each score interval Whites and Others will have the same number of

examinees. For each item, five tables are developed, one for each score level, of right/wrong by group. It is necessary to check that there are a minimum number of five examinees in each group who are expected to score correctly and about twenty examinees across groups who are expected to answer the item incorrectly. The most difficulty is encountered in setting the highest and lowest score levels. For instance, when investigating ethnic group differences often there are few high ability blacks who get the item wrong or low ability whites who answer the item correctly. The items which do not have enough examinees in each cell need to have the score levels adjusted (i.e., Lowest ability level total score changed from 1-15 to 1-20 and the other levels changed accordingly). In this study there were still five score levels for the MAT-J and MAT-K; however, this was not always possible on all GED items. Also, because the setting of score levels is arbitrary and determined by the required expected number of examinees in each cell, the size of the chi-square may change. In order to minimize the changes that could occur one tries to maintain the same number of examinees in each group (White and Others) at each level.

The chi-square value calculated for each item used the following formula (Alderman & Holland, 1980):

$$\chi^2 = \sum_i \sum_j \sum_k \frac{\left[n_{ijk} - \frac{(n_{ijk+})\,(n_{ijk})}{n_{i++}} \right]^2}{\frac{(n_{ij+})(n_{i+k})}{n_{i++}}} \qquad (1)$$

i = score level (i=1, 2 . . . 5).

j = group (j=1, 2 for White, Other).

k = item response (k=0 incorrect, 1 correct).

n_{ijk} = frequency of examinees within i^{th} score level from the j^{th} group with response k.

n_{ij+} = total frequency of examinees within i^{th} score level from j^{th} group regardless of item response.

n_{i+k} = total frequency of examinees within i^{th} score level with item response k, regardless of group.

n_{i++} = total frequency of examinees within the i^{th} score level, regardless of group or item response.

The degree of freedom for each chi-square is equal to $i(j-1)(k-1)$, where i represents the score level, j the number of groups and k the number of possible item

responses. Each score level can be assessed by omitting the summation over levels and using the individual score level values as a chi-square statistic with (j-1)(k-1) degrees of freedom.

The n_{ijk} represents the observed frequency of item performance by score level and group. The expected frequency ($\hat{n}_{ijk}$) of item performance for the same score level and groups is given by the subtrahend of the numerator and the denominator.

$$\hat{n}_{ijk} = \frac{(n_{ij+})(n_{i+k})}{n_{i++}} = n_{ij+}\frac{(n_{i+k})}{n_{i++}} \qquad (2)$$

The n_{ij+} is the total number of examinees from a particular group within a given score level. The proportion of examinees in the total sample from the same score level with response k is $\frac{n_{i+k}}{n_{i++}}$, which is independent of group membership. The expected frequency of item performance for a single cell is the product of these two terms.

The chi-square index given in equation (1) tests the hypothesis of no item bias, i.e. a triple interaction between ability level, ethnicity and response.

In this study the full chi-square index was treated in two ways. First, for each item the actual chi-square value for each score level was summed to yield a continuous chi-square(CS). Secondly, because most studies on item bias classify items as biased or not biased according to the method used and the criteria of that method or that definition, a dichotomous chi-square (DCS) was defined. If the continuous chi-square(CS) is significant at the .05 level then the dichotomous chi-square (DCS)=1; DCS=0 if CS not significant at the .05 level (df=i(j-1)(k-1) where i=score level, j=groups, and k=possible response, i.e. 5(2-1)(2-1)). An item where the DCS equals one can be considered an item on which item bias was demonstrated, while an item classified as zero did not demonstrate item bias. It must be noted that as sample size increases so does the chi-square values. Therefore, on the GED which has a smaller sample, fewer CS will reach significance.

As stated previously, each item was divided into five score intervals based upon total subtest score in order to control for differences in ability. In many instances, it was not possible to have an equal number of examinees in each score interval, but by inspecting the distribution of total score for each group and for those answering the item correctly or incorrectly, adjustments

were possible so that the number of score levels could be maintained. Also, with some alterations the minimum expected frequency in each cell necessary to insure the validity of the chi-square test could be obtained.

On the GED, however, because the sample size was much smaller, it was not always possible to maintain the five score intervals and fulfill the expected frequency required in some cells. On the 79 such items, some of which occurred on all the subtests, an alternative procedure was employed. Since the continuous chi-square does not use a significance procedure, the best combination of five levels was added for the CS on those items. When the closest to five score levels met the validity requirement for the minimum expected frequency in each cell, all five chi-squares were summed, including those which did not meet the requirement. The dichotomous chi-square is defined according to whether the CS is significant at the .05 level. Therefore, for the DCS fewer score levels were used, until the minimum expected frequency was obtained. In some instances this could only be done with two score levels. The chi-squares from these score levels were added and defined as 1 if CS significant at .05 level; zero otherwise. The degrees of freedom were calculated according to the number of score intervals used ($df=i(k-1)(j-1)$).

The Four Sampling Procedures

Much support exists in the literature on item bias for the use of samples which are matched on some criterion or contain equal numbers for the groups under investigation. In this study, data were collected on three sampling methods for the same population in order to assess the differences obtained from the results in each. This information is presented in Table 2. These three samples were used for four sampling methods. The total samples (T) included any examinees on whom ethnic information was available and who had completed at least half the subtest. These Total samples contained disproportionate numbers of Whites and Others in the ratio of approximately three White examinees to one Other. For the Random (R) sample, examinees were selected randomly from the total White population to equal the total number of examinees in the Other sample. The Matched (M) samples also contained the same number of White and Other examinees, but for this sample the White examinees were selected randomly from each total subscale score to match the distribution of Others total subscale scores. For the MAT-J and MAT-K, the Random sample used the same examinees for all subtests whereas for the GED each subtest had a different group of examinees. A different

Table 2

Number of Examinees in Each Sampling Procedure

Test	White	Other
MAT-J		
Total(T)	7620	1755
Random(R)	1755	1755
Matched(M)	1755	1755
Matched-signed(MS)	1755	1755
MAT-K		
Total(T)	7056	1874
Random(R)	1874	1874
Match(M)	1874	1874
Matched-signed(MS)	1874	1874
GED		
Total (T)		
Reading	1047	335
Math	1025	330
Science	1046	330
Random(R)		
Reading	335	335
Math	330	330
Science	330	330
Matched(M)		
Reading	335	335
Math	330	330
Science	330	330
Matched-signed(MS)		
Reading	335	335
Math	330	330
Science	330	330

sample of White examinees was used for each subtest for the matched sample because each subtest had different distributions of total scores.

In addition to these three continuous chi-squares and three dichotomous chi-squares, an additional analysis was done with the matched sample. Since for each item at the different score intervals the group of examinees whose expected frequency of correct answers is not always the same, it is possible that at one ability level the Whites observed frequency of correct scores is higher than expected, while at a different ability level Others may actually score correctly more times than expected. The literature describes this type of item as demonstrating "differential validity" (Scheuneman, 1976) which means that at one level the item is biased against one group, but at a different score level there is bias against another group.

In order to account for items with this pattern, the data from the Matched sample were examined to yield an additional CS and DCS. To accomplish this the actual chi-squares for each item were inspected at each score level to see whether the White or Others performed better than expected. If the White group had a larger observed frequency of correct responses than expected, the

chi-square was given a plus which meant that the item was biased against Others. If the Others obtained more correct answers than was expected the chi-square was given a minus: the item was biased against Whites. The signed values for each score interval, generally five, were then summed to yield a continuous chi-square for that item. Again, the dichotomous chi-square was defined as one if the continuous value was significant at the .05 level; zero otherwise. The level of significance although appropriate for the Total, Random and Matched DCS is not appropriate for the MS. Since each chi-square at each score level has had a sign attached to it, it is no longer the same figure. However, for purposes of consistency in this study, the MS-DCS has been defined in the same way as the other sampling procedures. Obviously, because the MS continuous chi-square contains both negative and positive values the MS values are often smaller, with fewer chi-squares reaching significance. This is especially true when the chi-squares at each score level are large but compensating.

Comparisons of groups then are four-fold: Total, Random, Matched and Matched-signed(MS). Figure 1 represents the eight dependent variables calculated for each subtest on each test to make twenty-four dependent

Figure 1

Dependent Variables

	Item Bias Measure	
Sample	**Continuous Chi-Square (CS)**	**Dichotomous Chi-Square (DCS)**
Total (T)	TCS	TDCS
Random (R)	RCS	RDCS
Matched (M)	MCS	MDCS
Matched-signed (MS)	MSCS	MSDCS

variables in all. Each column depicts the full chi-square index as either a continuous or dichotomous measure. Each row stands for the sampling method on which the bias measure was computed. The matched-signed row actually uses the same group of examinees as the Matched sampling method but, as explained, calculates the continuous chi-square somewhat differently to exclude bias against Whites.

ITEM PROPERTIES

At the preliminary stage many item characteristics were examined to determine which ones, as a set of properties, would predict item bias. Stepwise regression analysis was used to determine the most feasible set of item properties. On some analyses as many as twelve independent variables were entered. For instance, several variables were examined with respect to an item's placement on the page such as top half/bottom half, right side/left side, and quadrant 1, 2, 3 or 4. Item location within the test as a whole demonstrated a stronger relationship with the bias measure so that the others were eliminated. Differences in the presentation of the item stem were tested with regard to the number of sentences in the item compared to complete/incomplete sentences as well as the number of words in the stem. It was found that for the Reading

test the complete/incomplete stem showed a higher correlation with the bias measures, while for the Math and Science subtests the number of words in the stem had a higher correlation.

Other variables were considered in different ways. Reference to any group was scaled variously: reference to any group/no reference to any group; reference to minority group/no reference to minority group; reference to majority group/reference to minority group/no reference. Also several possibilities were examined as to the similarity of an item to a practice item. First, was the item exactly like the practice item in question and response format, or was the item question the same as the practice item, or was the item response format the same as the practice item? Secondly, were the item and the practice item exactly the same or not the same? Third, was the item similar to the practice item? For example, in math, were the item and practice item computation problems although not necessarily both division?

Many other variables were also considered but were rejected. The "correct answer" was evaluated in several ways: the position for the correct answer (A,B,C,D), the number of words in the response options until the correct answer, and whether the correct answer

was the longest or shortest response option. Differences in the length of the response options were examined as a categorical variable (The response options were the same or different in length.) and as a continuous variable which counted the number of words between the longest and shortest option. The presence of clues, either grammatical or numerical, in the item stem was another variable which did not correlate with the bias measures. Also, for the response format and for the question format, negative words, qualifiers and combination answers were tried first as separate categories before putting them together.

For each subtest the same set of seven predictors were employed. It was decided that seven was the maximum number of predictors possible because the items of each subtest, which ranged from 50 on the Math subtest to 55 on the Reading and Science subtests, must be thought of as the subjects and thus are a limiting factor. It was important to use the same predictors for each of the eight bias measures within each subtest so that comparisons could be made across the chi-square measures. The following method was used to choose the common set of seven characteristics. First, eight regression analyses were conducted within a given subtest for each bias measure. In each analysis up to twelve

characteristics were examined at once. Using as a stopping rule a statistically significant ($p<.05$) increase in the multiple correlation, the best seven variables were identified. Secondly, it was found that a common set of four or five item characteristics were selected in six or more of the eight sets. These common variables were then chosen. Thirdly, various candidates for the remaining three or four variables (i.e. to constitute a final set of seven variable) were identified. These three or four choices for the additional variables were considered, and those combinations which yielded the highest multiple correlations across the eight bias measures were chosen, thus defining the final set of seven variables. It was decided to not use less than seven variables because it was not possible to identify a common set without sacrificing the size of the multiple correlation coefficient.

To clarify this procedure consider the Math subtest as an example. The seven characteristics which yielded the highest squared multiple correlation for each of the eight chi-square measure are given in Table 3 under the column headed "R^2" over the squared multiple correlation for the seven best predictors. For each chi-square measure, the set of characteristics differs.

Table 3

Selection of Common Set of Seven Characteristics for Math

Chi-Square Measure	Set of seven variables yielding highest Squared Multiple Correlation(1)							R^2(2)	Final R^2 for 7 common variables * x_1 x_2 x_3 x_4 x_5 x_6 x_7
TCS	Pp	Ral	Qf	Rf	Rl	Twr	Math	.37	.35
TDCS	Ref	Prac	Stmwd	Ral	Rf	Rl	In	.32	.30
RCS	Prac	Stmwd	Qf	Rf	Rl	In	Math	.48	.42
RDCS	Prac	Qf	Rf	Twr	Caw	In	Stmwd	.36	.30
MCS	Ref	Ral	Qf	Rf	Rl	In	Math	.38	.35
MDCS	Ref	Prac	Qf	Rf	Rl	Caw	Math	.25	.22
MSCS	Prac	Stmd	Ral	Qf	Rl	Caw	Math	.35	.32
MSDCS	Ref	Prac	Ral	Rf	Rl	In	Math	.38	.28

(1) Chosen using a stepwise (maximum R^2) regression analysis.
(2) R^2 corresponding to the seven variables chosen by the stepwise procedure.
(3) R^2 for the "common" set of seven variables, denoted by asterisk below.

Key to characteristics

Pp		Placement of question on top or bottom half of page
*Ral	x_3	Difference in length of response alternatives
*Qf	x_4	Question format
*Rf	x_5	Response format
*Rl	x_6	Arrangement of response options
Twr		Total words in response options
*Math	x_7	Word problem or computation question
Ref		Reference to any group
*Prac	x_1	Similarity to practice item
*Stmwd	x_2	Number of words or numbers in question
In		Item number in subtest
Caw		Words in correct answer

When a tally was made of the number of times each characteristic was listed, it was found that five characteristics (Response format, Response list, Question format, Type of problem and Similarity to practice item) were included in six or more equations, and that Difference in response option length, Item location, Reference to any group and Stemwords were chosen in four or more equations. In this case the former set of five characteristics was chosen and then the latter set of four were examined for all eight bias measures to identify which characteristics would yield the highest seven variable multiple correlation across the eight bias measures. The final set of variables included Difference in response option length and Stemwords. The squared multiple correlations for this common set of variables are listed in Table 3 under the heading "Final R^2." While it is clear that the "Final R^2" values are lower than the values exhibited in the "R^2" column, in order to meaningfully compare results over the eight chi-square bias measures a common set of predictors was necessary. The type of results presented in Table 3 for the Math subtest, were also found for the Reading and Science subtests.

For the Reading test four categorical and three continuous variables were selected, defined and scaled as follows:

1) Reference to any group

Reference to any group indicates that the item contains any word in the question or response alternatives which specifies any group, or uses a pronoun which specifies any group such as "he" or "she". This includes any visual material which relates to the item.
0= no reference
1= reference

2) Similarity to practice item

The practice item is the same as the item in type and format for the question and response alternatives. For instance, if the practice item requires inferential reasoning and the question calls for factual information, the item is not the same as the practice item.
0= same as practice item
1= different from practice item

3) Item stem

0= one or more complete sentences
1= item stem not a complete sentence

4) The difference in response alternative length

The number of letters between the shortest and the longest response alternative including spaces was counted.

5) Question format

The question format is stated positively without any negative words, and without any qualifiers. This includes words such as POSSIBLY, PROBABLY, MAYBE, OPPOSITE or IF/THEN, EITHER/OR, LEAST, DIFFERENT FROM or negative contractions
0= clear, positive question
1= negative or ambiguous question

6) Item location Number of item in test.

7) Passage length Number of lines in each reading passage upon which the question is based.

In order to clarify the scaling procedure a practice item and one item from the GED Practice Test, Reading, Form A are presented here followed by the scaling of the item characteristics.

Practice item:

Blow, bugle, blow, set the wild echoes flying,
Blow, bugle; answer, echoes, dying, dying, dying.

What happened to the sound the bugle makes?

(1) It becomes annoying to the listener.
(2) It fades as it is repeated.
(3) It is joined by other sounds.
(4) It becomes more joyful.
(5) It becomes an answer to the listener's question.

Item:

Questions 5-6 refer to the following poem.

What Is Poetry?

What is poetry? Who knows?
Not the rose, but the scent of the rose;

Not the sky, but the light of the sky;

Not the fly, but the gleam of the fly;

Not the sea, but the sound of the sea;

Not myself, but what makes me
See, hear, feel something that prose
Cannot, and what it is, who knows?

5. What is the main idea of the poem?

(1) Poetry should be enjoyed by everyone.
(2) Poetry cannot be defined in words.
(3) Poetry is more useful than prose.
(4) Poetry is composed by rhyming words.
(5) The best poems are those about nature.

Scaling of item 5:

1) Reference to any group = 0. There is no reference to any group, majority or minority.

2) Similarity to practice item = 1. Question 5 asks for the main idea of the poem, whereas the practice item wants the examinee to interpret the poem.

3) Item stem = 0. Item stem is a complete sentence.

4) The difference in response alternative length = 5. Option (3), the shortest option, equals 32 letters and spaces while option (5) equals 37 letters and spaces.

5) Question format = 0. Clear, unambiguous question without qualifiers or negative words.

6) Item location = 5. This is the fifth item in this subtest.

7) Passage length = 9. The selection is nine printed lines long which includes the title.

The properties of the Science subtest which were analyzed comprised two categorical and five continuous variables.

1) Stem words. Number of words in question not counting "a" or "an".

2) Difference in response options. Number of letters or numbers different between shortest and longest response option.

3) Response alternative placement

The options should be listed vertically in a column with the response alternatives placed one under the other.
0= options listed vertically
1= options horizontal or horizontal and vertical

4) Total words in response alternatives

Total number of words in response options, not counting "a" or "an".

5) Correct answer words.

Total number of words in correct response option, not counting "a" or "an".

6) Subtest item location. Number of item in test.

7) Visual material

The presentation of any visual material referred to by a question or as part of the question or options. This includes flow charts, graphs, diagrams or any material not written as a sentence or part of a sentence.
0= no visual material
1= visual material

An example of the scaling method for Science follows with only an item included because the similarity to practice item was not an independent variable for this subtest. This item is from the Practice Tests of GED, Science, Form A.

Item:

6. Fundamental life processes that take place in most animal cells include all of the following EXCEPT

(1) obtaining and using energy
(2) giving off oxygen
(3) eliminating waste

(4) reproducing themselves
(5) getting food

Scaling of item 6:

1) Stem words = 16. The total number of words in the question.

2) Difference in response option length = 14. Option (4) is the shortest with twelve letters and spaces, whereas option (1) is the longest with 26 letters and spaces.

3) Response alternative placement = 0. All options are listed vertically.

4) Total words in response alternative =13. The total number of words in all the response alternatives.

5) Correct answer words = 3. Number of words in correct answer.

6) Subtest item location = 6. This is the sixth item on subtest.

7) Visual material = 0. The question or response options does not contain any visual material.

The Math subtest was scaled on five categorical and two continuous variables.

1) Similarity to practice item

The practice item is the same as the item in type and format for the question and response alternatives. For instance, if the practice item requires inferential reasoning and the question calls for factual information, the item is not the same as the practice item. A word problem which requires two computational steps is not considered the same as a practice item which needs only one step to find the answer.
0= same as practice item
1= different from practice item

2) Stem Words.

Number of words or numbers in question not counting "a" or "an." An underlined blank or blank box was considered a word. [☐]

3) Difference in response options.

Number of letters or numbers different between the shortest and longest response option.

4) Question Format

The question format is stated positively without any negative words, and without any qualifiers. This includes words such as POSSIBLY, PROBABLY, MAYBE, OPPOSITE or IF/THEN, EITHER/OR, LEAST, DIFFERENT FROM or negative contractions.
0= clear, positive question
1= negative or ambiguous question

5) Response alternative format

The response alternative requires that one best answer is required. There should be no alternatives which ask for combination answers such as ALL OF THE ABOVE, NONE OF THE ABOVE, AND/OR, a and c; a,b and c, etc. There should not be any negatives included in the choices including negative contractions. There should not be any qualifiers such as ON or NEAR.
0= Best answer
1= combination or negative

6) Response alternative placement

The options should be listed vertically in a column with the response alternatives placed one under the other.
0= options listed vertically
1= options horizontal or horizontal and vertical

7) Type of problem

Word problems are considered any item where there are written words. Computation problems contain only numbers.
0= computation problem
1= word problem

To illustrate the Math scaling procedures, a practice item and one item from the GED Practice Test, Mathematics, Form A are listed.

Practice item: What is the cost of 5 quarts of milk at 45 cents a quart?

(1) $1.80
(2) $2.05
(3) $2.25
(4) $2.35
(5) $2.70

Item: 9. Of the 29,580 adults arrested for major crimes in one state during one year, 10,514 were convicted and 5,239 of these were eventually imprisoned. How many of those convicted were not imprisoned?

(1) 24,341 (2) 19,066 (3) 13,827
(4) 5,326 (5) 5,275

Scaling of item 9:

1) Similarity to practice item = 0. The practice item is one-step word problem and so is the item.

2) Stem words = 32. The total number of words and numbers in the item stem.

3) Difference in response option length = 1. Options (4) and (5) are five numbers, the shortest options and options (1), (2) and (3) are 6 spaces long.

4) Question format = 1. The negative word NOT appears in the question.

5) Response alternative format = 0. Only one answer required without any qualifying words.

6) Response alternative placement = 1. Options are listed across.

7) Type of problem = 1. The question is a word problem.

The use of the chi-square index rests upon the assumption that the test is unidimensional. To verify this assumption, factor analyses were performed on the three subtests from the MAT, Form J. For each subtest, Reading, Math and Science a separate analysis was undertaken which used the unrotated principal factor procedure. Tetrachoric correlations of each item with every other item of each subtest were computed and entered as the input matrix. For the Reading subtest one factor was produced which accounted for 91.6 per cent of the common factor variance; the second factor accounted for 8.4 per cent. There was one major factor which accounted for 83.3 per cent of the common variance on the Math subtest; the second factor accounted for 16.6 per cent of the common variance. In the Science subtest, 90.6 per cent of the common variance was accounted for by the first factor; and 9.3 per cent was accounted for by the second factor. These figures support the assumption that the three subtests from which the prediction equations were calculated are unidimensional. In each case one major factor accounted for most of the variance.

CHAPTER V: RESULTS

EXAMINEES

For the two groups under investigation in this study, White and Others, the means and standard deviations on each subtest for the samples appear in Table 4. The means and standard deviations for Others are only presented once because the Total group of Others was used in the Random and Matched samples. Also, the Matched-Signed(MS) sampling method used the same group of examinees as the Matched sample.

It can be seen that in every instance the White group had a higher mean score on each subtest than the Others. The difference between the two groups is largest for the Total sample and smallest for the matched sample. The White group also has larger standard deviations, except on the GED Reading and Math.

CHI-SQUARE INDEX

The means over items for each subtest and test of the continuous chi-square(CS) measures for this study are given in Figure 2. Several trends are apparent. The pattern of means across subtests for the various sampling procedures appears quite similar. The Total groups have the largest means for each test across subtests with the exception of the GED Science. Conversely, all the means

Table 4

Means(M), Standard Deviations(SD) and Group Size(n) of Test Scores

Test	Sample											
	TOTAL			WHITE RANDOM			WHITE MATCHED			OTHER T,R,M		
MAT-J	M	SD	n	M	SD	n	M	SD	n	M	SD	n
Reading	35.0	11.8	7620	34.8	11.7	1755	26.4	11.2	1755	26.3	11.3	1755
Math	28.8	9.0	7620	28.7	8.8	1755	22.3	7.7	1755	22.3	7.7	1755
Science	31.1	9.4	7620	31.0	9.4	1755	23.7	7.8	1755	23.6	7.8	1755
MAT-K												
Reading	34.3	12.0	7056	34.1	12.2	1874	26.0	11.4	1874	25.9	11.4	1874
Math	28.5	9.1	7056	28.3	9.2	1874	22.0	8.0	1874	22.0	8.0	1874
Science	30.7	9.2	7056	30.6	9.2	1874	23.3	7.6	1874	23.3	7.6	1874
GED												
Reading	27.0	7.0	1047	27.2	7.2	335	22.2	7.1	335	22.1	7.1	335
Math	26.8	6.9	1025	26.8	7.1	330	22.4	7.2	330	22.2	7.5	330
Science	34.0	9.0	1046	34.1	8.6	330	28.2	8.8	330	27.8	8.8	330

are smallest for the matched-signed (MS) measure. While it cannot be seen on this figure, these small means for the MS group are accompanied by the largest, or second largest, standard deviation. The MAT-K has the highest mean values on all the subtests and across sampling procedures with the exception of the MS groups and Total-Science.

The proportions and frequencies of the significant DCS shown in Table 5, are the same as the percentage of items where the dichotomous chi-square value equals one. For these items, the continuous chi-square was significant at the .05 level, and can be considered as demonstrating item bias. After each percentage the number of items is given in order to compare the number of items across samples of the same test. While it is informative to compare the MS sample to the other samples it must be remembered that the summing of the signed values does not control for the overall level of significance.

In all cases the Total group has the highest percentage of biased items and the matched-signed group the smallest percentage. The GED has less than half the percentage of biased items than the MAT-J and MAT-K. The range of biased items varied from one item on the GED Reading MS sampling to thirty-five items on the MAT-K

Table 5

Proportions and Frequencies of Significant Chi-Square Values

	TESTS		
	MAT-J	MAT-K	GED
Reading	**N=55** [a]	**N=55**	**N=40**
Total [b]	.45(25)[c]	.64(35)	.15(6)
Random	.31(17)	.53(29)	.15(6)
Matched	.35(19)	.53(29)	.08(3)
Matched-signed	.07(4)	.15(8)	.03(1)
Math	**N=50**	**N=50**	**N=50**
Total	.52(27)	.70(35)	.16(8)
Random	.30(15)	.60(30)	.08(4)
Matched	.38(19)	.58(29)	.08(4)
Matched-signed	.16(8)	.24(12)	.04(2)
Science	**N=55**	**N=55**	**N=60**
Total	.44(24)	.44(24)	.13(8)
Random	.27(15)	.38(21)	.13(8)
Matched	.29(16)	.40(22)	.05(3)
Matched-signed	.09(5)	.16(9)	.03(2)

[a] N = Number of subtest items

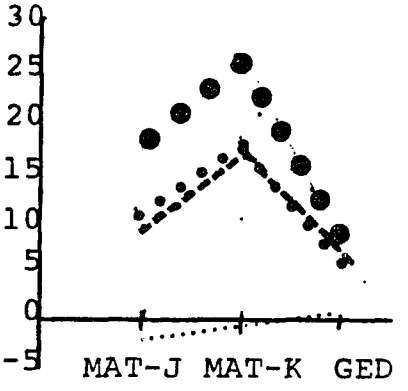
[b] The number of examinees in each sample is given in Table 2

[c] Number of items where DCS is significant

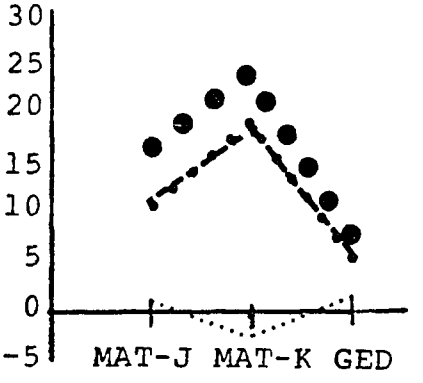
Figure 2

Means of Continuous Chi-Squares by Subtest

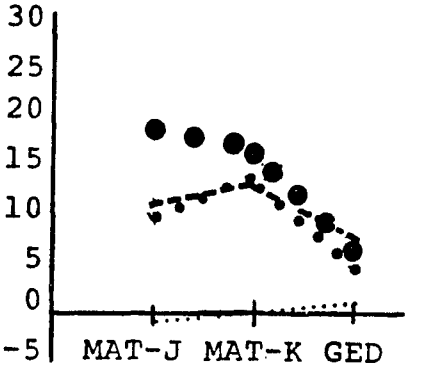
Reading



Math



Science



Key
● ● = Total
---- = Random
• • • = Matched
..... = Matched-signed

Reading Total group and MAT-K Math Total group. Seventy per cent of the items on the MAT-K Math for the Total sampling method were biased. Interestingly, the two forms of the Metropolitan Achievement Test do not have comparable percentages of items classified as demonstrating item bias, especially on the Reading and Math subtests. In fact even the pattern of biased items for the different sampling procedures within the Reading and Math subtests is different for Form J than for Form K. On all the subtests of the MAT-J and MAT-K and on the GED Math the Random and Matched sampling methods have nearly the same number of biased items.

A comparison of the Pearson product moment correlations of the eight chi-square values for each subtest for the MAT-J, MAT-K, and GED demonstrate almost identical patterns. All the correlations are significant at $p<.05$ except for some of the Matched-signed(MS) chi-squares(CS and DCS). On each test the range was from one to five nonsignificant correlations for the MS continuous chi-squares. On the MAT-K Math and GED Math the correlation of the MSDCS with the RDCS was nonsignificant. The eight different chi-square measures appear to be identifying the same construct.

INDEPENDENT VARIABLES

A different set of independent variables, or item characteristics, was selected for each subtest. As previously noted, however, within a subtest the same seven characteristics were used for the final analyses. Each set included a combination of categorical and continuous variables. The means and standard deviations for the independent variables for each test and subtest are presented in Table 6. The categorical variables used zero/one dummy coding. For instance, on the Reading subtest one variable concerns the reference to, or no reference to, any group. The item would be in category 1 if there was any reference to any group, minority or majority. The mean for the categorical variables therefore is the same as the percentage of items in category one. In this case fifty-eight percent of the items on the MAT-J contained a reference to some group. On the GED Reading only about one-fourth the items referred to some group.

On the Reading subtest about half the items on all three tests were similar to the practice item. The item stem was written as an incomplete sentence in 85, 96 and 70 percent of the questions on the MAT-J, MAT-K, GED respectively. This meant that only two Reading items on the MAT-K were complete sentences. A greater percentage

Table 6

Means and Standard Deviations of Independent Variables

Variables	TESTS		
	MAT-J	MAT-K	GED
Reading	N=55	N=55	N=55
Reference x_1*	.58(.50)	.55(.50)	.23(.42)
Practice item x_2*	.47(.50)	.49(.50)	.53(.42)
Stem length x_3*	.85(.36)	.96(.19)	.70(.46)
Response length x_4	11.73(6.97)	11.96(7.82)	14.58(14.47)
Question format x_5*	.13(.34)	.13(.34)	.24(.44)
Item location x_6	28.00(16.02)	28.00(16.02)	20.50(11.69)
Passage length x_7*	23.42(6.41)	26.16(6.47)	27.43(6.41)
Math	N=50	N=50	N=50
Pratice item x_1*	.78(.42)	.92(.27)	.86(.35)
Stemwords x_2	9.04(7.09)	10.80(9.46)	22.58(11.71)
Response length x_3	2.22(2.71)	1.94(2.74)	1.96(5.94)
Question format x_4*	.06(.24)	.10(.30)	.16(.37)
Response format x_5*	.38(.49)	.36(.46)	.06(.24)
Response list x_6*	.40(.49)	.46(.50)	.68(.47)
Word problem x_7*	.52(.50)	.51(.50)	.10(.30)
Science	N=55	N=55	N=60
Stemwords x_1	12.53(6.72)	13.55(7.34)	23.85(10.55)
Response length x_2	9.95(8.52)	11.93(10.90)	21.89(16.93)
Response list x_3*	.58(.50)	.56(.50)	.15(.36)
Total words x_4	15.02(12.54)	15.58(11.48)	27.69(18.15)
Words of correct answer x_5	3.89(3.25)	4.02(3.23)	5.62(3.75)
Item location x_6	28.00(16.02)	28.00(16.02)	30.50(17.46)
Visual material x_7*	.44(.50)	.78(.50)	.18(.39)

Note:
* Categorical variables
Numbers in parentheses are standard deviations

of questions on the MAT-J and MAT-K than on the GED are written without any qualifying or negative words. The difference in the number of letters between the longest and the shortest response option appears to be similar for all the tests, although the range on the GED is much larger.

The MAT-J and MAT-K have a similar ordering of variables on the Math subtest, while the GED varies considerably. Few items were similar to the practice item on any math test, but especially on the MAT-K. It can be seen that the MAT-J and MAT-K contain about fifty per cent of each type of question but that the GED contained only ten per cent word problems. Also, the questions on the GED on the average possessed many more words than the two forms of the MAT. However, many more response options on the MAT-J and MAT-K were combination answers, such as NONE OF THE ABOVE, or NOT GIVEN.

On the Science subtest, again the GED questions are much longer on the average with more than twice the number of words than on the MAT-J or MAT-K. This is also reflected in the total words in all the response options where the range of words for the GED is greater as well as the mean number of words in the response options. Interestingly, this discrepancy does not exist for the correct answer where the range and number of words in all

the tests are similar. Many more Science items on the MAT-K include visual material than on the MAT-J or GED. The arrangement of response options is entirely vertical in less than half the MAT-J and MAT-K tests; but 85 per cent of the GED items have vertical response lists.

REGRESSION EQUATIONS

The regession weights both within each test and across sampling procedures vary considerably. These weights have been examined in three ways. First, the raw regression weights for each subtest can be found in Table 7. Next, the standardized regression weights are presented in Table 8. Finally, Kendall's coefficient of Concordance has been calculated for each subtest from the rankings of the standardized regression weights for each chi-square measure.

It is of interest to identify the variables which are important in predicting the bias measure. If one defines importance in terms of the standardized regression weights, Table 8 presents this information for analysis. For the Reading subtest variable x_7, passage length, receives the largest positive standardized weight for the Total, Random and Matched continuous chi-square as well as the Total and Random dichotomous chi-square. For the Matched CS, the highest posistive standardized

Table 7

Raw Regression Coefficients for MAT-J

	TCS	TDCS	RCS	RDCS	MCS	MDCS	MSCS	MDCS
Reading								
Intercept	42.46	.73	1.27	.43	18.69	.78	6.36	.49
Reference x_1	1.09	.03	.67	.08	2.03	.19	-3.32	.91
Practice item x_2	1.98	.15	-1.32	-.07	-3.56	.05	2.88	.06
Stem length x_3	-16.05	-.32	-5.23	-.40	-6.30	-.43	2.67	-.14
Response length x_4	-1.10	-.02	-.47	-.03	-.54	-.03	-.06	-.01
Question format x_5	4.78	.37	1.22	.09	-4.69	-.27	3.35	-.14
Item location x_6	-.40	-.02	-.15	-.01	-.09	.00	.01	.00
Passage length x_7	.46	.02	.37	.03	.25	.01	-.39	.00
Math								
Intercept	-8.05	.27	1.50	.06	2.36	-.11	-4.48	.22
Practice item x_1	.05	.30	3.51	.29	-1.49	-.09	-3.38	-.20
Stemwords x_2	.03	.04	.02	.01	1.09	.01	-.31	.01
Response length x_3	1.24	.02	.48	.01	.91	.01	1.35	.02
Question format x_4	31.06	.30	14.39	.60	15.05	.62	19.75	.52
Response format x_5	20.47	.43	9.92	.35	14.30	.56	6.14	.35
Response list x_6	14.80	.35	.34	-.16	8.13	.28	12.47	.37
Word problems x_7	10.68	-.31	2.18	-.05	4.73	.20	12.87	.23
Science								
Intercept	23.81	.74	9.26	.08	12.60	.14	5.04	.27
Stemwords x_1	.49	.01	.21	.01	.04	.01	.25	.01
Response length x_2	.71	.02	.32	.02	.42	.02	-.23	.00
Response list x_3	-10.10	-.36	-.13	.01	-1.21	-.14	-4.38	-.17
Total words x_4	2.19	-.04	-.53	-.03	.24	-.01	-.07	.01
Words: correct ans. x_5	6.67	.06	1.58	.08	-1.88	-.03	.78	-.05
Item location x_6	-.28	-.01	-.09	.00	-.08	.00	-.26	.00
Visual material x_7	3.23	.15	-.67	.12	-1.33	.00	2.18	.05

Table 8

Standardized Regression Coefficients for MAT-J

	TCS	TDCS	RCS	RDCS	MCS	MDCS	MSCS	MDCS
Reading								
Reference x_1	.03	.03	.05	.09	.13	.19	-.17	.02
Practice item x_2	-.06	.15	-.10	-.07	-.22	-.05	.15	.11
Stem length x_3	-.33	-.23	-.27	-.31	-.28	-.32	.10	-.19
Response length x_4	-.44	-.30	-.47	-.37	-.48	-.36	-.04	-.28
Question format x_5	.09	.24	.06	.06	-.20	-.19	.12	-.18
Item location x_6	-.37	-.54	-.35	-.41	-.17	-.14	.01	-.17
Passage length x_7	.17	.29	.34	.42	.20	.15	-.26	-.10
Math								
Practice item x_1	.00	.25	.17	.26	-.06	-.07	-.38	-.23
Stemwords x_2	.01	.59	.01	.15	.07	.10	-.15	.11
Response length x_3	.19	.09	.15	-.07	.24	.08	.25	.16
Question format x_4	.42	.14	.39	.31	.34	.30	.32	.34
Response format x_5	.56	.41	.55	.37	.67	.56	.20	.46
Response list x_6	.41	.34	.02	-.17	.38	.28	.42	.50
Word problems x_7	.30	-.31	.12	-.06	.23	.20	.44	.32
Science								
Stemwords x_1	.18	.19	.16	.19	.03	.19	.14	.11
Response length x_2	.33	.32	.30	.37	.40	.43	-.16	.10
Response list x_3	-.28	-.36	.01	.01	-.07	-.16	-.18	-.30
Total Words x_4	-1.15	-.92	-.74	-.85	.32	-.19	-.07	.31
Words: correct ans. x_5	1.20	.40	.57	.60	-.68	-.19	.20	-.53
Item location x_6	-.25	-.23	-.15	-.11	-.14	.06	-.33	-.23
Visual materials x_7	.09	.15	-.04	.14	-.07	.00	.09	.09

regression coefficient is found for x_1, reference to any group and x_7, the length of the passage, is the second largest positive value. Similarity to the practice item, x_2, has the largest positive standardized weight for the MS samples, both CS and DCS.

More than half the standardized regression weights are negative and while some are smaller than -.10, most are larger. In five of eight analyses, x_4, the difference in letters and spaces between the shortest and longest response option, has the largest negative weight as well as being the highest in magnitude for the TCS, RCS, MCS, MDCS and MSDCS. For the TDCS and RDCS, x_6, the item location is the largest negative weight, and again, the largest weight. For the Total and Random measures when variable x_4 was the largest negative weight, variable x_6 was next highest and if x_6 was highest, x_4 was second in size.

The standardized regression weights for the Math subtests in Table 8 indicate that the independent variable x_5, representing response format, results in the largest or second largest increase in the measure of item bias for all but the MSCS index. For that sampling method the largest positive standardized regression coefficient is x_7, word problem or computation, with the second highest coefficient found for x_6, response list arrangement.

The number of words in the item stem, x_2, contributes most to the TDCS regression equation while x_6, response list arrangement, has the largest positive standardized regression weight for the MSDCS. The biggest negative standardized regression coefficients can be found for x_1, similarity to practice item, on both Matched and MS measures, although small in magnitude for the Matched measures. TCS and RCS do not contain negative standardized weights. The variables x_7, word problem or computation, is negative only for TDCS and RDCS while most of the other analyses have a sizeable positive weight for this variable.

The standardized regression coefficients for Science in Table 8 present a confusing picture for Matched and MS sampling methods, but it can be seen that the Total and Random CS and DCS are almost identical. Words in the correct answer, x_5, had the largest positive coefficient for these four measures and the MSCS. Also this variable had the largest negative weight as well as being the largest in magnitude for MSDCS and MCS. The largest negative standardized regession weight is x_4, total words in all the response options, for TCS, TDCS, RCS and RDCS, while it was the largest positive weight for MSDCS.

Another way of comparing the standardized regression weights across chi-square measures and sampling procedures is by calculating Kendall's Coefficient of Concordance. This coefficient is one value, on a scale of zero to one, that describes the extent to which a set of k rank orderings of N things tend to agree. In this case there are eight chi-square measures (k) for each of seven independant variables (N). If the standardized regression weights for each chi-square index are ranked in absolute value from one for the largest weight to seven for the smallest, the similarity of ranking across measures can be assessed. The coefficient of concordance, W, computes the variability among the sum of the ranks for each independent variable.

$$W = \frac{\text{variance of rank sums}}{\text{maximum possible variance of rank sums}}$$

or according to Siegel (1956) p. 235

$$W = \frac{s}{\frac{1}{12} k^2 (N^3-N)} \qquad (3)$$

where s = sum of the observed deviations from the mean of the sum of the ranking (R_j) that is, $s = \sum \left(R_j - \frac{\sum R_j}{N} \right)^2$

It should be noted that W is linearly related to the average Spearman rank correlation coefficients taken over all groups. This relationship is expressed as

$$\text{average } r_s = \frac{kW - 1}{k-1} \qquad (4)$$

Also the significance of W can be found by evaluating the probability of an s of some magnitude. The null hypothesis states that the R_j sets of rankings are independent.

For the Reading subtest W equals .34, which suggests that the agreement of the rankings of the independent variables for the eight chi-square measures is weak. However, the s associated with this value of W is significant at the .01 level of significance. For the Math subtest W was calculated as .42. This value is higher than the Reading test but is still only moderate and also is significant at $p<.01$. On the Science subtest $W = .52$, significant at $p<.01$. This is the highest coefficient for the three subtests which means that there is more agreement of ranking of variables between the various chi-square measures on this subtest than Reading or Math.

MULTIPLE CORRELATIONS

The multiple correlations of the seven item characteristics for each subtest of the MAT-J with the different chi-square values for each sampling procedure are listed in Table 9. Several multiple correlations are relatively high in the range, though nonsignificant, because the number of items in each test was small. The number of items range from 50 items on the Math subtest to 55 items on the Reading and Science subtests. In this study the number of items can be thought of as subjects so that in determining the multiple correlations, the small number of items was a limiting factor.

On the Reading subtest the multiple correlations varied from .41 MSDCS to .61 for TDCS. Three CS correlations were significant at $p<.01$ while two DCS correlations were significant.

All the multiple correlations of the CS were significant on the Math subtest, and only the matched sample was not significant for the DCS. The multiple correlations on the Math subtest were higher, as a group, than the Reading and Science subtests and also the values for the CS were higher than the DCS for each sample.

The set of item properties from the Science subtest had the smallest multiple correlations with the

Table 9

Multiple Correlations of Chi-Squares with Item Properties for the MAT-J

	CS	DCS
Reading		
T	.59**	.61**
R	.55**	.55*
M	.55**	.43
MS	.46	.41
Math		
T	.59**	.52*
R	.65**	.55*
M	.59**	.47
MS	.57*	.53*
Science		
T	.56**	.47
R	.36	.38
M	.38	.39
MS	.45	.39

* p< .05
** p< .01

measures of item bias. For the Total sample and the Matched-signed sampling method the continuous chi-square was higher. While the Random and Matched sampling groups are similar for the CS and DCS.

Differences were found in the multiple correlations for the four sampling methods. All but one value is significant for the Total group and the Random has four significant values. The range of values is greatest for the Random sample: the Random-Math subtest has the largest multiple correlation while the Random-Science subtest has the smallest value.

CROSS VALIDATION

The regression weights from the MAT-J for the three sets of item characteristics, one set for each subtest, were cross-validated on the Reading, Math and Science subtests of the MAT-K and the GED. Each new subtest was scaled on the appropriate set of item characteristics, then the set of regression weights from MAT-J for that sampling method and that subtest were applied in order to predict a continuous and dichotomous chi-square for each item. The observed chi-squares were also calculated for each item of each sampling procedure on each subtest. The correlation of the predicated values with the observed values was computed and can be found in Table 10.

Table 10

Correlation of Predicted Chi-Square with Observed Chi-Square by Subtest

	MAT-K		GED	
	CS	DCS	CS	DCS
Reading				
T	.34**	.15	-.13	-.05
R	.25	.15	-.05	-.04
M	.18	-.29*	-.07	-.07
MS	.32**	.41**	-.16	-.15
Math				
T	.14	.28*	-.15	-.22
R	-.03	-.03	-.11	-.16
M	-.02	-.04	-.12	-.17
MS	.44**	.33*	-.07	-.13
Science				
T	.20	-.03	-.02	-.26*
R	.29	.00	.17	.05
M	.27*	.02	-.06	.00
MS	.19	.06	-.05	.35**

* $p < .05$

** $p < .01$

In general, most of the regression equations did not show cross-validations that were significantly different from zero. In fact, only about twenty per cent of the correlations were significant, which was ten out of a possible forty-eight correlations. Further, the results where the regression equations were applied to a new test, i.e. GED, were far less significant, with only one significant positive correlation, then when the equations were applied to a parallel form of the same test, i.e. MAT-K.

On the MAT-K Reading subtest four correlations are significant: two CS for the Total and MS groups and two DCS, for the Matched and MS sampling methods. The Math subtest has the widest range of correlations from -.04 for MDCS to .44 for the MSCS. Both MS chi-squares are significant as is the TDCS. The MAT-K Science subtest contains two significant CS values but all the DCS are zero or close to zero. For the GED, the only significant correlation was found for the Science subtest: .35 for MSDCS. All the Reading and the Math subtests have negative correlations.

The correlations, by sampling method, for the chi-square values predicted by the weights from the MAT-J

with the observed values from the MAT-K and GED, show that the matched-signed procedure demonstrated the most significant values. The Random sample contains the weakest correlations.

CHAPTER VI: DISCUSSION AND RECOMMENDATION

The primary purpose of this study was to determine if a set of observable item charactertics could be defined which would be predictive of a statistical index of item bias. It was hoped that the prediction equations derived from a multiple regression analyses based on the Metropolitan Achievement Test, 1978, Form J (MAT-J), and the full chi-square measure as an index of item bias could be generalized to other samples, such as the Metropolitan Achievement Test, 1978, Form K (MAT-K) and to other tests, such as the General Educational Development Test (GED). It is of great practical value if such equations could be developed and used in item analyses. Previous research has focused on identifying biased items post hoc, which has not benefitted test developers in the construction of bias-free or bias-reduced items. While it is necessary to be able to detect the presence of biased items within a test, the problem can never be corrected unless these items can be changed or altered before the test is administered.

The goal of this study was to investigate whether a strategy could be defined which could, a priori, i.e. at the time of test construction, minimize item bias. To determine how well such a strategy succeeded the various elements of this study can be

analysed. First, was it possible to define a set of observable item properties which would correlate with the bias measure, the full chi-square index? Secondly, could one develop regression equations based upon these multiple correlations? Third, how well did these regression weights predict biased items on other tests?

Item properties were precisely defined and scaled in three content areas: reading, math and science. The 1978 standardization sample of the MAT-J was used to examine the test results of two groups, White and Others. The full chi-square measure was the bias measure calculated for each item of the three subtests: Reading, Math and Science. A continuous chi-square(CS) was computed for each item as well as a dichotomous chi-square(DCS). The DCS equalled one if the CS was significant at $p<.05$, zero otherwise. Four sampling procedures were defined to compare the results. First, the total sample of each group, White and Other was included in the Total group(T). Second, a random sample of Whites was selected to equal the total number of Others(R). Third, a sample of Whites was selected for each subtest to match the total subscale score distribution of the total Other group(M). Finally, the chi-square index for these same groups matched on each subtest score was calculated differently. At each score

level it was determined which group had a higher expected frequency of correct responses than observed frequency. If the Whites had a higher expected frequency of correct responses than the observed frequency, that indicated bias against Whites, and was given a negative sign. If the Other group had a higher expected frequency than was observed the sign of the chi-square was positive and indicated bias against Others. When these chi-squares were summed for each item across score levels the negative values dropped out; or stated another way, only bias against Others was included. This sampling procedure was called Matched-signed(MS). As mentioned earlier the MSDCS has been used for comparison purposes only. Because signs have been attached to the chi-square values at each score level, the significance level is altered.

Consequently, twenty-four regression equations were developed on the MAT-J: eight each for Reading, Math and Science. The two chi-square measures (CS and DCS) were computed for each of the four sampling procedures (T, R, M, MS). The equations were used for cross-validation on the Reading, Math and Science subtests of the MAT-K and GED.

The results indicate that it was possible to define item characteristics which would, as a set,

correlate with the measures of item bias. Many item properties were considered and rejected because the correlations with the item bias measure were too low. It must be noted, however, that this study used national standardized tests where rigorous item analyses and refinement had already occurred. If an item pool had been analyzed, more than likely, some of the item properties which were not included may have shown higher correlations with the bias measures. Actually, it could be argued that some of the item characteristics which were selected should not have been found to correlate with the bias measures on these tests since most manuals for item writers list as pitfalls to avoid in constructing test items such characteristics as combination answers (e.g. A and B, A and C, etc.) or the sterotypic reference to some group (e.g. Ellen baked a cake using 4 cups of flour, etc.).

Unfortunately, the same set of item properties could not be defined in all three content areas: Reading, Math and Science. Again, proper screening of the item pools may have eliminated items which refer to some group, or which were very unlike the practice item. It is possible that this elimination is more successful in some content areas such as Math and Science, where these variables were not included as contrasted with the

Reading test where they were considered. Also, since the variable selected which were decidedly specific to a subtest (i.e. passage length, word problem or computation, and presence of visual material) were important to the multiple correlations it may be that it is not fruitful to try to find item characteristics that are generalizable to all content areas. After all, reading questions refer to skills and material important to learning to read, whereas multiplication and division problems test mathematical ability at a certain level. Therefore, it does not seem unreasonable to find that a different set of item properties is important in different content areas.

Significant multiple correlations of the item properties with the chi-square measures were demonstrated across subtests. The results on the Math subtest, where seven multiple correlations are significant, out of a possible eight, are especially heartening. More than half the multiple correlations on the Reading subtest are also significant. And, even though only one multiple correlation was significant on the Science subtest, several other multiple correlations are relatively high.

Twenty-four prediction equations were developed based upon the multiple correlations. Regression equations could be calculated which were able to predict

bias. The results of the regression analyses suggest the following recommendations for item writers. These recommendations are based on the size and signs of the standardized regression weights. First, on Reading tests all passages may need to be the same length, or not vary greatly in length. Incomplete item stems seem to predict a smaller bias index when all other characteristics are held constant. One possible explanation is that an incomplete stem may be shorter which may be the influencing factor. For Math and Science, the fewer the number of stem words in the questions, the smaller the bias measure, all other characteristics being the same. Also, for these two subtests the inclusion of modifying words in the question or answers, as well as the presence of combination options appears to increase the bias measure with the other properties held constant. It seems relevant to the Math subtest, but not for the Science, that response options should be listed vertically only. This format may lessen the confusion with numbers on the Math subtest, but may not be necessary on the Science, where the options are often words. On the Science subtest, when the other variables were held constant, the bias index increased as the number of words in the response options decreased. However, the bias measure increased as the number of words in the correct answer increased. This seeming

paradox may indicate that the more words combined in the response options may present more information from which to select the correct answer; but that the correct answer is clearer with fewer words.

The results from the cross-validation indicated that the prediction equations worked better for the MAT-K, a different form of the same test with the same age group than for the GED, an entirely different test administered to a different age group. The pattern of significant correlations between the predicted bias measures and the observed bias measures presented in Table 10 is similar to the pattern of multiple correlations of item properties with the bias measures for the MAT-K only. Significant correlations can be found across subtests, and there are more significant correlations on the Reading and Math subtests than on the Science. The correlations for the GED were only positive and significant for the MSDCS for Science. A cross-validation could have been undertaken on a subsample of the MAT-J which most likely would have demonstrated higher correlations than those obtained with the MAT-K and GED. This idea was rejected because that cross-validation would have no real life generalizability to other forms and other tests.

One might conjecture as to why the cross validation analyses produced generally poor results. First, the ability level of the examinees who took the GED may have differed significantly from those examinees who took the MAT. Another reason for these results may be that the GED was not a very biased test, which can be seen from Figure 1 and Table 5. The low cross validities can be partly explained in terms of this restriction in range for the bias measure.

The usefulness of the regression weights seems limited to tests which are administered to similar, if not the same, populations. Even though the content areas are the same for the MAT-J, MAT-K and GED tests, the MAT tests are administered to the junior high school population while the GED is designed for older teenagers and adults. It may be found that test constructors could develop a set of equations on one form of a test which then could be applied to other forms of the same test. Also, the prediction equations were more accurate for the Reading and Math subtests than the Science subtest. Content specifications are more precise for these two areas than for Science and consequently reading and math tests are probably more comparable across tests.

Most research to date on the problem of item bias has classified items as biased or not biased as opposed to measuring item bias on a continuum. Not only do few definitions of this dichotomy across methods agree, but also when the same method is used in different studies the definition varies. Two approaches are generally followed: an a priori number of items is selected as representing "most" bias (Nungester, 1977) or else a statistic such as 1.5 standard deviations above the mean of the distribution of the item bias measure is used (Burrill, 1980). In this study, no prior decision was made as to how many items would be considered biased, but instead the continuous CS was tested for significance. Those items which were significant were classified as biased. All analyses were calculated for both the continuous and dichotomous measure in order to ascertain which measure was the best predictor. In Table 9 eight significant multiple correlations can be found for the CS and five multiple correlations were significant for the DCS. However, on the cross-validation, Table 10, five significant correlations each were found for the DCS and CS. There appears to be no evidence which favors the preference of the continuous(CS) or dichotomous(DCS) chi-square index. If the DCS is defined as a significant CS though, consideration must be given to chi-square values which may be inflated due to large sample size.

In one sense the eight bias measures are quite similar, yet in another sense dissimilar. The high correlations among the eight indices indicate that the same items are being selected as most or least biased by all the measures and sampling procedures. On the other hand, the differences between the eight measures on the multiple correlations and standardized regression coefficients suggest that for the different measures, different item characteristics have a different influence. However, still another important consideration in choosing a sampling procedure or a chi-square index is the ease of computation. The Total and Random samples were the simplest to select and on which to compute a chi-square index; while the Matched and Matched-signed sampling methods were more difficult. The Total sampling method identified many more items as biased than any of the other sampling methods whereas the Random and Matched procedures identify a similar, smaller, number in the majority of cases. Since the MS sampling method eliminates bias in one direction, in this study against Whites, the fewest items were identified. If the ease of computation is the major factor in determining one sampling method to use, then the Random sampling procedure would seem a reasonable choice.

An important difference between the Total, Random and Matched sampling methods and the MS procedure is that bias in any direction is included in the first three procedures, but that bias against Whites is subtracted from the MS procedure. This means that a decision must be made about whether the concern is item bias in any direction, or whether one wishes to identify only item bias against minority groups. Other indices could also have been included such as a signed index which measures bias against Whites. Further research should be undertaken with a Random-signed sample in order to compare it with the MS sampling method. Again, the MS and a Random-signed procedure may identify the same items as biased so that only one index would be necessary if minority bias is the concern.

The use of the full chi-square procedure posed several computational problems, yet is theoretically promising. The greatest difficulty encountered concerned the development of score levels. This was especially true for the smaller GED sample. Scheuneman (1981) and others (Rudner, Getson and Knight 1980; Shepard, Camilli and Averill, 1980) have stated that the full chi-square method could be employed with as few as 200 examinees in a group. Even though the GED analysis exceeded this number there was not always a sufficient number of

examinees to derive five score levels and have the minimum expected frequency requirement met. The Random, Matched and Matched-signed samples were at times difficult to complete with over 330 or more examinees in each group. On very easy or very hard items there were not enough examinees who had different subscale scores to put in each score level. This made the number who answered the question correctly on an easy item, extremely disproportionate to the number of examinees who missed the item. The distributions of right/wrong by group for these items, even for the Matched sample, was sometimes quite dissimilar. Some items, again mostly on the GED, needed many manual adjustments in order to have the required expected frequency in all the cells. While this is not difficult, it is tedious and time consuming. On the GED, 79 items could not be divided into five score intervals, and for the DCS fewer score intervals were used.

The function of the score interval is to a establish some control of ability as a confounding factor in assessing item bias. The score levels can be thought of as discrete and increasing stairs compared to the smooth curve of the ICC. In the full chi-square procedure the probability of a correct or incorrect response within a score level is the same for all groups.

Consequently, if one score level has many more examinees than the others, or if the range of subscale scores is wider than the range of subscale scores in another score interval, it is difficult to justify the assumption that the probability of answering correctly has the same meaning across score levels. Also, it must be noted that the chi-square value changes when the score intervals are changed. The use of the total subscale score as the ability measure has been criticized because this score contains items which may then be classified as biased. However, the score levels include a rather wide range of possible scores within a level not just one value so that the inclusion of biased items is somewhat minimized.

The ease of computation, once the score levels have been derived, and the theoretical simplicity of the method, are reasons why the full chi-square method should continue in use as a bias measure. One would hope that a more universally agreed upon definition for dichotomizing this index would be developed. The dichotomy established in this study seems a reasonable statistical approach.

A review of the major findings of this study demonstrates that first and most important, it was possible to define a set of item properties for the content areas of Reading, Math and Science which would

correlate highly with the bias measures. Second, it was possible to develop regression equations which would predict item bias. Third, the prediction equations which were developed generalized moderately with a different form of the same test (MAT-K) given to the same age examinees, but did not generalize to a different, although similar, test (GED) given to a different age group. Methodologically, the two chi-square indices computed for the four sampling procedures indicate that there was much agreement among these eight measures except for the Matched-signed continuous chi-square.

What implications are there for further research? It is important to note that this is one of the only research studies to investigate item bias a priori and try to develop a procedure to use before the final test form is constructed. Consequently, a natural outgrowth of this work would be analyses of item pools in the same manner as the test items here. Each item can be scaled on the relevant variables for that subtest and the items which predict the highest bias measure can then be changed or deleted. Then the remainder of the item pool including the changed items could be readministered, and a new bias measure for each item calculated. The results would demonstrate whether the same items are still biased, and if so to what degree,

and whether items which previously were not biased are now identified as biased. In other words, what effect will altering the item characteristics have on the bias measures.

Additional cross-validation studies might further clarify the efficacy of the prediction equations. For example, a cross-validiation study which would compare different, but similar, tests such as the MAT and the Stanford Achievement Test administered to the same age group may find more significant correlations than those obtained from the GED. Finally, it would be informative to redo the analyses in this study using fewer item characteristics. While this would mean a reduction in the multiple correlations, the prediction equations may prove more useful.

The present findings can provide the foundation from which to identify more clearly the influence of observable and measurable item characteristics on item bias. The results are of practical importance since they provide a basis for removing bias in the test construction stage rather than after the test has been administered. Although the results were not completely satisfactory, the evidence reported in this study should provide some guidelines for item writers and test developers before a final test form is completed. Most

of the characteristics identified here can be modified easily without altering the essential content of the item. The practical implications of these findings should not only be the basis of further study in measurement research, but also can be implemented immediately by test publishers.

VII: BIBLIOGRAPHY

Alderman, D. L. & Holland, P. W. Item performance across native language groups on the Test of English as a Foreign Language (TOEFL Research Report 9) Princeton, N. J.: Educational Testing Service, 1980.

Angoff, W. H. A technique for the investigation of cultural differences. Paper presented at the meeting of the American Psychological Association, Honolulu, September 1972. (ERIC Document Reproduction Service No. Ed 069 686).

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

Angoff, W. H. The use of difficulty and discrimination indices for detecting item bias. In. R. A. Berk (Ed.), Handbook of methods for detecting test bias. Baltimore, Md.: Johns Hopkins University Press, 1982.

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

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

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick, Statistical theories of mental test scores. Reading, Massachusetts: Addison-Wesley, 1968.

Blew, E. and Ishizuka, T. College board item bias study of the Scholastic Aptitute Test and the Test of Standard Written English -- Form XSA4/E7. (ETS SR 78-62). Princeton, N. J.: Educational Testing Service, 1978.

Blew, E. and Stern, J. College board item bias study of the Scholastic Aptitude Test and the Test of Standard Written English -- Form XSH5/E8. (ETS SR 79-37). Princeton, N. J.: Educational Testing Service, 1979.

Bode, R. K. Comparison of pretest and reanalysis results of an item bias study. Paper presented at the meeting of the American Educational Research Association, Los Angeles, Calif., April 1981. (ERIC Document Reproduction Service No. ED 208 036).

Breland, H. M. An investigation of cross-cultural stability in mental test scores. Paper presented at American Educational Research Association, Chicago, Illinois, April 1974.

Burrill, L. E. A comparative investigation into the identification of ethnic bias in items assessing current educational status (Doctoral dissertation, Fordham University, 1981) Dissertation Abstracts International, 1981, 42A, 1110. (Univerity Microfilms No. 81-19762).

Cardall, C. & Coffman, W. E. A method for comparing the performance of different groups on the items in a test (ETS RB 64-61). Princeton, N. J.: Educational Testing Service, November 1964.

Cleary, J. A. Test bias: Prediction of grades of Negro and white students in integrated colleges. Journal of Educational Measurement, 1968, 5, 115-124.

Cleary, J. A. & Hilton, T. L. An investigation of item bias. Educational and Psychological Measurement, 1968, 28, 61-75.

Coffman, W. E. Sex differences in response to items in an aptitude test. Eighteenth Yearbook, National Council on Measurement Education, 1961, 117-124.

Cole, N. S. Bias in selection. Journal of Educational Measurement, 1968, 5, 115-124.

Crowder, C. R. An investigation of item bias occurring at different ability levels for Anglo and Mexican-American students. Paper presented at the meeting of the American Educational Research Association, San Francisco, Calif., April 1979. (ERIC Document Reproduction Service No. ED 174 677).

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

Devine, P. J. and Raju, N. S. Extent of overlap among four item bias methods. Educational and Psychological Measurement, Winter 1982, 42, 1049-66.

Diamond, E. E. Item bias issues: Background, problems and where we are today. Paper presented at the meeting of the American Educational Research Association, Los Angeles, Calif., April, 1981 (ERIC Document Reproduction Service No. ED 200 631).

Donlon, T. F., Ekstrom, R. B., Lockheed, M. and Harris, A. Performance consequences of sex bias in the content of major achievement batteries. Princeton, N. J.: Educational Testing Service, 1977.

Douglas, J. B. Item bias test speededness and Rasch tests to fit. Paper presented at the meeting of the American Educational Research Association, Los Angeles, Calif., April 1981.

Dudycha, A. L. and Carpenter, J. B. Effects of item format on item discrimination and difficulty. Journal of Applied Psychology, 1973, 58, 116.

Durovic, J. J. Test bias: An objective definition for test items. Paper presented at the meeting of the Northeastern Educational Research Association, Ellenville, N. Y., October, 1975. (ERIC Document Reproduction Service No. ED 128 381).

Echternacht, G. A quick method for determining test bias. Educational and Psychological Measurement, 1974, 34, 271-280.

Einhorn, H. J. and Bass, A. R. Methodological considerations relevant to discrimination in employment testing. Psychological Bulletin, 1971, 75, 261-260.

Faggen-Steckler, J., McCarthy, K. A. & Tittle, C. K. A quantitiative method for measuring sex "bias" in standardized tests. Journal of Educational Measurement, 1974, 11, 151-161.

Fishbein, R. L. An investigation of the fairness of the items in a test battery. Paper presented at the meeting of the National Council on Measurement in Education, Washington, D. C., April 1975. (ERIC Document Reproduction Service No. ED 111 837).

Flaugher, R. L. The many definitions of test bias. American Psychologist, 1978, 33, 671-679.

Flaugher, R. L., Milton, R. S. and Myers, C. J. Item rearrangement under typical test conditions. Educational and Psychological Measurement, 1968, 28, 813-824.

Frary, R. B. and Giles, M. B. Multiple choice test bias due to answer strategy variation. Paper presented at annual meeting and American Educational Research Association, Boston, Mass., April, 1980.

Green, D. R. Racial and ethnic bias in test construction: Final Report. (Adapted from Final Report of U. S. O. E. Contract No. OEC-9-70-0058 [057]). Monterey, California: CTB/McGraw-Hill, 1971. (ERIC Document Reproduction Service No. ED 056 090).

Green, D. R. Reducing bias in achievement tests. Paper presented at the meeting of the American Educational Research Association, San Francisco, California April 1976. (ERIC Document Reproduction Service No. ED 126 126)

Green, D. R. and Draper, J. F. Exploratory studies of bias in achievement tests. Paper presented at the meeting of the American Psychological Association, Honolulu, September 1972. (ERIC Document Reproduction Service No. ED 070 794).

Gross, A. L. & Su, W. H. Defining a "fair" or "unbiased" selection model: A question of utilities. Journal of Educational Measurement, 1975, 60, 345-351.

Haebara, T. A. A method for investigating item bias using Birnbaum's three parameter logistic model, Iowa City, Iowa Testing Programs, University of Iowa, Number 25, December 1979.

Hoyt, C. Test reliability obtained by analysis of variance. Psychometrika, 1941, 6, 153-160.

Huck, S. W. and Bowers, N. D. Item difficulty level and sequence effects in multiple-choice achievement tests. Journal of Educational Measurement, 1972, 9, 105-111.

Hughes, H. H. and Trimble, W. E. The use of complex alternatives in multiple choice items. Educational and Psychological Measurement, 1965, 25, 117-126.

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

Huntley, R. M. & Plake, B. S. Effects of selected item-writing practices on test performance: can relevant grammatical clues result in flawed items. Paper presented at Amercian Educational Research Association, Boston, Mass., 1980 (ERIC Document Reproduction Service No. ED 189 115)

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

Ironson, G. H. and Craig, R. Item bias techniques where the amount of bias is varied and score differences between groups are present. University of South Florida, Tampa, Fla., 1982.

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

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

Kolen, M. J. and Hoover, H. D. The reliability of selected item bias procedures. Paper presented at the meeting of the American Educational Research Association, New York, N. Y., March, 1982.

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

Linn, R. L., Levine, Hastings & Wardrop. An investigation of item bias in a test of reading comprehension. (Tech Rep. 163) Urbana, Illinois: Illinois University, Center for the Study of Reading, March, 1980 (ERIC Document Reproduction Service No. ED 184 091).

Linn, R. L. & Harnisch, D. L. Interactions between item content and group membership on achievement test items. Journal of Educational Measurement, 1981, 18, 109-118.

Lloyd, B. W. Analysis of content-related item bias for Anglo and Hispanic students. Paper presented at meeting of American Educational Research Association, New York, N. Y., March, 1982.

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

Lord, F. M. Applications of item response theory to practical testing problems. Hillsdale, N. J.: Erlbaum, 1980.

Lord, F. M. & Novick, M. R. Statistical theories of mental test scores. Reading, Massachusetts: Addison-Wesley, 1968.

Marascuilo, L. & Slaughter, R. E. Statistical procedures for identifying possible sources of item bias based on χ^2 statistics. Journal of Educational Measurement, 1981, 18, 229-248.

Marco, G. L. Item characteristic curve solutions to three intractable testing problems. Journal of Educational Measurement, 1977, 14, 139-160.

McCarthy, K. Sex bias in tests of mathematical aptitude. (Doctoral dissertation, City University of N. Y., 1976) Dissertation Abstracts International, 1977, 36A, 7301. University Microfilms No. 76-11629).

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

Merz, W. R. A factor analysis of the Goodenough-Harris Drawing Test across four ethnic groups (Doctoral dissertation, The University of New Mexico, 1970). Dissertation Abstracts International, 1970, 31, 1627A. (University Microfilms No. 70-70, 714).

Merz, W. R. Estimating bias in test items utilizing principal components analysis and the general linear solution. Paper presented at the meeting of the American Educational Research Association, San Francisco, Calif., April 1976. (ERIC Document Reproduction Service No. ED 129 871).

Merz, W. R. & Grossen, N. An empirical investigation of six methods of examining test item bias. Paper presented at the meeting of the National Council on Measurement in Education, San Francisco, Calif., April 1979.

Nungester, R. J. An empirical examination of three models of item bias (Doctoral dissertation, The Florida State University, 1977). Dissertation Abstracts International, 1977, 38, 2626A. (University Microfilms No. 77-24789).

Oosterhof, A. C. Atash, M. N. and Lassiter, K. L. Facilitating identification of item bias through use of delta plots. Paper presented at meeting of American Educational Research Association, New York, N. Y., March, 1982.

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

Pine, S. M. Applications of item response theory to the problem of test bias. Washington, D. C.: Office of Naval Research, Personnel and Training Branch, 1976. (ERIC Document Reproduction Service No. ED 159 196).

Plake, B. S. & Hoover, H. D. A methodology for identifying biased achievement test items that removes the confounding in an item by groups interaction level. Paper presented at the meeting of the American Educational Research Association, Toronto, March 1978. (ERIC Document Reproduction Service No. ED 161 930).

Plake, B. S., Hoover, H. D. & Lloyd, B. An investigation of differential item performance by sex on the Iowa Test of Bias Skills. Paper presented at the meeting of the National Council on Measurement in Education, Toronto, March 1978. (ERIC Document Reproduction Service No. ED 161 933).

Pothoff, R. F. Statistical aspects of the problem of biases in psychological tests. (Institute of Statistics Mimeo Series No. 479). Chapel Hill: University of North Carolina, Department of Statistics, May 1966 (Second printing, February 1972).

Rasch, G. An item analysis which takes individual differences into account. British Journal of Mathematics and Statistical Psychology. 1966, 19, 49-59.

Roid, G. H. and Wendler, C. L. Item bias detection and item writing technology. Paper presented at meeting at American Educational Research Association, Montreal, Quebec, April, 1983.

Rudner, L. M. An approach to biased item identification using latent trait measurement theory. Paper presented at the meeting of the American Educational Research Association, New York, N. Y., April 1977. (ERIC Document Reproduction Service No. ED 137 337). (The original title contained the word methodology, rather than theory.)

Rudner, L. M. & Convey, J. J. An evaluation of select approaches for biased item identification. Paper presented at the meeting of the American Educational Research Association, Toronto, March 1978. (ERIC Document Reproduction Service No. ED 157 942).

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

Rudner, L. M., Getson, P. R. & Knight, D. L. Biased item detection techniques. Journal of Educational Statistics, 1980, 5, 213-233. (b)

Scheuneman, J. A new method of assessing bias in test items. Paper presented at the meeting of the American Educational Research Association, Washington, D. C., April 1975, (ERIC Document Reproduction Service No. ED 106 359)

Scheuneman, J. Validating a procedure for assessing bias in test items in the absence of an outside criterion. Paper presented at the meeting of the American Educational Research Association, San Francisco, Calif., April 1976. (ERIC Document Reproduction Service No. ED 129 853)

Scheuneman, J. Further considerations in the assessment of bias in test items. Paper presented at the meeting of the American Psychological Association, Toronto, August 1978.

Scheuneman, J. D. A response to Baker's criticism. Journal of Educational Measurement, 1981, 18, 63-66.

Scheuneman, J. D. A posteriori analyses of biased items. In R. A. Berk (Ed.), Handbook of methods for detecting test bias. Baltimore, Md.: Johns Hopkins University Press, 1982. (a)

Scheuneman, J. D. Item bias and test scores. Paper presented at meeting of National Council on Measurement in Education, New York, N. Y., March, 1982. (b)

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

Shepard, L., Camilli, G. & Averill, M. Comparison of six procedures for detecting test item bias using both internal and external ability criteria. Paper presented at the annual meeting of the National Council on Measurement in Education, Boston, Mass., 1980.

Siegel, S. Nonparmetic statistics for the behavioral sciences. New York: McGraw-Hill, 1956.

Sinnott, L. T. Differences in item performance across groups (ETS RR-80-19). Princeton, N. J.: Educational Testing Service, 1980.

Strassberg-Rosenberg, B. & Donlon, T. F. Content influences on sex difference in performance on aptitude tests. Paper presented at the meeting of the National Council on Measurement in Education, Washington, D. C., April 1975. (ERIC Document Reproduction Service No. 110 493).

Stricker, L. J. A new index of differential subgroup performance: Application to the GRE Aptitude Test (ETS Research Report, RR-81-13). Princeton, N. J.: Educational Testing Service, 1981.

Subkoviak, M. J., Mack, J. S. & Ironson, G. H. Item bias detection procedures: empirical validation. Paper presented at American Educational Research Association, Los Angeles, Calif., April 1981.

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

Thurstone, L. L. A method of scaling psychological and educational tests. Journal of Education and Psychology, 1925, 16, 433-451.

Tollefson, N. and Tripp, A. The effect of item format on item difficulty and item discrimination. Paper presented at meeting of American Educational Research Association, Montreal, Quebec, April, 1983.

Tucker, L. R. An inter-battery method of factor analysis. Psychometrika, 1958, 23, 111-136.

Urry, V. W. Ancillary estimators for the parameters of mental test models. Paper presented at the meeting of the American Psychological Association, Chicago, Ill., August 1975.

Veale, J. R. & Foreman, D. I. Cultural variation in criterion-referenced tests: A 'global' item analysis. Paper presented at the meeting of the American Educational Research Association, San Francisco, Calif., April 1976.

Whitely, S. E. & Dawis, R. V. The influence of test context on item difficulty. Educational and Psychological Measurement, 1976, 36, 329-337.

Wightman, L. E. Study of LSAT item performance for different groups. (ETS Draft for Committee Review) Princeton, N. J.: Educational Testing Service, 1979.

Wright, B. D., Mead, R. J. & Draba, R. E. Detecting and correcting test item bias with a logistic response model. Chicago: University of Chicago, Department of Education, Statistical Laboratory, Research Memorandum No. 22, October 1976.

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

Yen, W. M. and Herman, J. L. Using the nonfinishers responses to examine item bias. Research in Education, December, 1981. (ERIC Document Reproduction Service No. 205 578).

Yen, W. M. The extent, causes and importance of context effects on item parameters for two latent trait models. Journal of Educational Measurement, 1980, 17, 297-311.

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