This is to certify that the

dissertation entitled

COMPARING CHI-SQUARE AND LOG-LINEAR METHODS OF DETECTING DIFFERENTIAL ITEM PERFORMANCE ON A MINIMUM COMPETENCY TEST

presented by

MARTHA S. JONES

has been accepted towards fulfillment
of the requirements for

Ph.D. degree in EDUCATIONAL MEASUREMENT

William R Mehrens
Major professor

Date 8-9-88

COMPARING CHI-SQUARE AND LOG-LINEAR METHODS OF DETECTING DIFFERENTIAL ITEM PERFORMANCE ON A MINIMUM COMPETENCY TEST

By

Martha S. Jones

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Counseling, Educational Psychology, and
Special Education

1988

ABSTRACT

COMPARING CHI-SQUARE AND LOG-LINEAR METHODS
OF DETECTING DIFFERENTIAL ITEM PERFORMANCE
ON A MINIMUM COMPETENCY TEST

By

Martha S. Jones

This study was concerned with comparing three methods of detecting differential item performance. The methods were chosen for their suitability for tests conducted with a small number of students, with more than two ethnic or other groups, in a mastery test situation, or without the resources to employ item response theory methods. The methods studied were all based on contingency-table analysis: the Scheuneman chi-square, the full chi-square, and logit-linear analysis. In previous research, the first two methods have been judged the next best alternative to item response theory.

The test data were obtained in a regular administration of the Michigan Educational Assessment Program, a mandated statewide minimum competency test. The fourth grade test included 75 reading items and 84 mathematics items. The final sample of subjects consisted of 3695 fourth grade students in public schools; they were classified by ethnicity (White, Black, or Hispanic), sex, and language dominance (English or other). They were divided into five groups, approximately

equal in size, for reading and for mathematics according to their total test score in that area in order to control for the effect of ability on item performance.

The Scheuneman and full chi-square methods were applied to each item twice, once for ethnicity and once for sex. The logit-linear method required only one use per item, because it could handle multiple independent variables and their interactions. The methods demonstrated moderately high correlations in identifying differentially performing items but could not be considered interchangeable.

In regard to item content, there was no consistent pattern in the identification of items. Although the reading items were somewhat more likely to favor Whites and females, no substantive generalizations could be drawn about the types of item content most likely to show differential performance.

The logit-linear method has theoretical promise as a way of examining several variables and their interactions at once in order to achieve a more complete understanding of the factors affecting item performance. For purposes of test construction, however, the present state of development suggests using the full chi-square method.

In loving memory

ALAN KIRKES MANCHESTER

1897 - 1983

MARY ELIZABETH ONDERDONK MANCHESTER

1900 - 1988

ACKNOWLEDGEMENTS

Conducting a dissertation research project often seems to be a solitary experience. In reality, many people enabled me to pursue a doctoral degree and to achieve my academic goals.

From the time he taught a first-term graduate student the basics of educational measurement, Bill Mehrens has been a sincere and concerned educator. His counsel as my academic advisor and my dissertation chair has been invaluable.

I owe Susan Phillips my gratitude for her encouragement and guidance. I also wish to thank Robert Floden and Neal Schmitt for their challenging standards and thoughtful suggestions.

Ed Roeber of the Michigan Department of Education was instrumental in obtaining the data for my analysis and in supplying pertinent insights. Both he and the participating school districts were invariably cordial and helpful.

The completion of my doctoral studies was greatly aided by the material and psychological resources provided by my friends Martha Seaman, Mary Fielding, and Josephine Hussain. Janet Vredevoogd deserves a special mention for her tireless assistance with computer programming and other areas. I also wish to thank my colleagues Karen Lassiter, Cindy Sopko, and Lucille Dungan for their advice and support during the writing of the dissertation.

Finally, no words can express my appreciation for my parents, Darrell and Beth Jones, for my husband, Charles Nicholson, and for our daughter, Amy Nicholson. Their love, patience, and understanding made my success possible.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1

INTRODUCTION

The present concern about promoting excellence in our country's educational systems has given new emphasis to the use of educational testing for the purposes of selection, certification, and evaluation. Several national reports have called for new or expanded testing programs, and many state and local districts are planning responses. "High-stakes" tests are taking on increasing importance for grade-to-grade promotion, high school graduation, college admission, and professional certification.

At the same time, educators and the public are committed to providing equity as well as excellence. Official policy and social opinion hold that educational opportunities should be open to all citizens and that no one should be held back by factors and circumstances beyond personal control. Educational testing and assessment should likewise be equitable processes for all examinees, even though the outcomes may differ.

One factor in the "fair testing" debate is the issue of bias in testing. Claims about the existence of bias and unfairness in testing, dating back at least to the Lippman-Terman debate of 1922-23 (Lippman, 1986; Terman, 1986), often occur in the context of more general controversy about the role and impact of testing (Cronbach, 1975). The discussion intensified in the late 1960's and 1970s (e.g., Williams, 1971) and has broadened to involve the judicial system in cases such as Larry P. v. Riles, Debra P. v. Turlington, and Golden Rule v. Washburn. Some critics of testing have gone so far as to label

certain tests "educational genocide" (M.E.R. Hoover, 1984). The critics generally base their position on grounds either of face invalidity associated with perceived sociocultural differences or of disparate outcomes in average scores or success rates (Jensen, 1980).

Measurement experts take a different approach to the definition of bias. For instance, Scheuneman (1982a) illustrates the concept of bias with this model:

$$X = \theta + \beta + \delta$$ where X = observed score, θ = true score, β = bias component, δ = measurement error

Because theoretically measurement error has a mean of zero, then if there were no bias component, the observed scores would be unbiased estimates of the true score. If β exists and has a non-zero mean, the observed scores will not be accurate estimates of the true scores.

In this context bias can be thought of as a source of invalidity in a score. Bias operates as an unexpected factor, an unwanted dimension, that impedes measurement of the ability represented by the true score.

Bias, loosely defined for the moment as an irrelevant effect of demographic variables on measured performance, can be found at several levels of measurement. (Actually demographic characteristics such as sex or ethnicity may simply be proxies for variables such as opportunity to learn the performance being measured.) The first level is the use of a total test score for selection or admissions decisions when the correlation of that score with the desired criterion is

affected improperly by group membership. A related concept is that of response pattern; because a total test score is composed of many right and wrong answers, there may be considerable variability in the pattern of answers, and hence the interpretation, of a given score. Another level is the individual item, which may be biased if demographic factors wrongly affect performance. The most discrete level of measurement is that of distractor analysis, in which the choice of a particular answer option may turn out to be correlated with personal characteristics. Methods have been developed and tested for detecting bias at all these levels.

Differential Item Performance

The focus of the research presented here is on two methods for discovering bias at the level of individual items. First, however, it should be pointed out that what any item bias technique uncovers is not really "bias" per se. The techniques can only identify a discrepancy or difference in the behavior of items, and thus "differential item performance" or "differential item functioning" have come to be the terms preferred by researchers in the field. Actual bias, deliberate or unintended, may be the cause of the differential performance, but its presence can only be inferred. The item does not necessarily have an intrinsic bias; it is simply different from the other items in the test. As Shepard, Camilli, and Averill (1981) so clearly state:

> Bias cannot be identified in an isolated test item. Test questions designed to measure the same construct must be studied together; bias is discovered when an item does not fit the pattern established by others in the set. Thus, the bias assessed by these techniques is 'anomaly in a context of other items'; it is not bias in the sense of unfairness. (pp.3-4)

This study adopts the idea of differential item performance, although for brevity the term "bias" is usually employed. The most obvious source of a difference in item performance between demographic groups may be a true difference in the underlying ability being measured. For this reason most item bias detection techniques attempt to control for overall difference in ability by grouping on total test score, matching on an external criterion, or using an IRT true score. Then a biased item becomes one on which examinees of equal ability but from different demographic groups score differently.

The many types of item bias detection techniques, as well as several studies comparing their accuracy and utility, will be discussed in greater detail in Chapter 2. Many empirical studies, however, share two common and disconcerting findings that warrant consideration at this point. The first is that it is difficult to derive any general characteristics of item content, context, or format that account for the differential performance observed. Concrete explanations have been offered in a few cases (Scheuneman, 1979); for example, negatively phrased statements and Roman numerals seem to cause problems for some examinees in certain ethnic groups. Usually, though, the flagged items seem quite similar to others on the same test that show no discrepancy. Thus bias research has provided little guidance to test developers and educators who need to understand why the items behave as they do and who hope to avoid such problems in the future.

Perhaps it should not be surprising that the reasons for differential item performance are hard to discern. After all, there has not been much theoretical work on what qualities make an item difficult or discriminating generally. Some researchers are beginning

to pursue this issue (Scheueneman, personal communication, April 6, 1988), and others have conducted bias studies in which item characteristics are deliberately manipulated in order to test hypotheses about the causes of differential performance (Schmeiser, 1982; Scheuneman, 1987). Most empirical research to date, as discussed in Chapter 3, has been carried out on nationally published and standardized tests. Such tests usually review items repeatedly for content validity and statistical quality, eliminating those with obvious flaws or poor functioning. Items found acceptable after such extensive review may not represent the full range of possibilities for differential item performance.

The second concern is that the number of items identified as biased may be only a small percentage of the total and have little influence on the ranking of the affected examinees. Removal of the flagged items and rescoring the test may not affect score differences between groups enough to be worthwhile or may adversely affect test reliability and validity (Frary & Zimmerman, 1983). One counter-argument to this position is that in some uses of tests -- for example, a mastery test with a fixed passing score -- even one flawed item can have a serious impact on the number of examinees passing. Were this a "high-stakes" test such as one required for high school graduation, the consequences of a few flawed items could be severe.

A more general response to this issue focuses on the professional standards of test developers and users. Their desire for quality should be thorough and consistent. A misspelled word in a reading passage might not have any untoward effect on the examinees, but a reputable test developer would still correct the error. Similarly, an

item known to favor one group over another for reasons unrelated to the ability being measured does not enhance a test's reliability and validity. As one publisher writes (ETS, 1980):

> It is not futile to continue attempts at ferreting out any real or potential sources of unfairness. The guiding rule for testing experts is that they must strive to see that the assessment procedure itself does nothing that in any way could make matters worse. Because of the critical importance of testing, and the high visibility of the results of assessment, they must be vigilant in the pursuit of whatever sources of unfairness they might discover. The assessment must be free from any distortion it is possible to detect, whether or not there is an impact on the mean differences. (p. 12)

Purpose of the Study

This study explores the utility of two types of methods suitable for detecting item bias on a criterion-referenced test constructed to measure mastery of instructional objectives. One method is a chi-square technique; the other uses log-linear analysis. The study applies these methods to the Michigan Educational Assessment Program (MEAP) fourth grade reading and mathematics tests to identify differences in performance among demographic groups.

Why were these methods chosen? The research design examines the relationship of nominal variables - sex, language dominance, and ethnicity - to a dichotomous item response, controlling for ability. The usual statistic to test for independence in this classic nonparametric situation is the chi-square. Several comparative studies, as discussed in the subsequent review of literature, have rated the chi-square methods more highly than all others except three-parameter item response theory. The latter, however, requires large sample sizes and is expensive and complex to use. There

continues to be a need for methods such as chi-square which can be used on smaller samples (e.g. pilot testing) and are easier to explain to test users.

As described in the literature review, the chi-square approach has two versions. The full (Camilli) X2 uses all responses; the Scheuneman C2 uses only correct responses. Although the X2 is recommended for theoretical reasons discussed in Chapter 3 (Baker, 1981; Marascuilo & Slaughter, 1981), empirical comparisons of the X2 and C2 have found reasonable agreement between the two versions. These studies, however, used typical norm-referenced data. The MEAP is an objective-referenced test with many easy items and a negatively skewed distribution. Scheueneman (1977) suggests that the C2 is especially well suited to a test with such characteristics because it is not inflated by a small number of incorrect responses.

Question 1: How well do the chi-square methods, X2 and C2, agree in measuring differential item performance on a minimum competency test?

The chi-square methods employ one independent variable. To examine more than one variable, e.g. both ethnicity and sex, a corresponding number of chi-square indices must be computed for each item. Furthermore, there is no easy way to test possible interactions. Therefore the log-linear method, a "multidimensional chi-square," seems promising as a way to examine all variables of interest simultaneously. It has been recommended for item bias research (Mellenbergh, 1982; Marascuilo & Slaughter, 1981) but has seldom been given practical application.

Question 2: How well do chi-square and log-linear methods for detecting differential item performance agree?

Although the focus of this study is methodological, it does use real data from real examinees on a real test. Because the test is objective-referenced, the intended classification and association of items is clear. Thus some secondary questions with a substantive focus can be considered.

Question 3: Is there evidence of differential item performance by ethnic group on the MEAP Grade 4 reading and mathematics tests? If so, are there any interpretable patterns?

Question 4: Is there evidence of differential item performance by language group on the MEAP Grade 4 reading and mathematics tests? If so, are there any interpretable patterns?

Question 5: Is there evidence of differential item performance by sex on the MEAP Grade 4 reading and mathematics test? If so, are there any interpretable patterns?

CHAPTER 2

REVIEW OF RESEARCH

This review will describe various techniques used to detect differential item performance, summarize the studies that compare them, and discuss the considerations important in selecting an appropriate technique.

Early Developments

The earliest psychometricians, including Binet and Stern, were concerned with equity in testing. Binet tried to select items for his intelligence tests that measured changes in mental ability rather than social class. The 1937 Stanford-Binet revision achieved equal score distributions for both sexes by discarding or counterbalancing the items with the greatest discrepancies in performance between boys and girls. In the same era test developers were pursuing the idea of construct validity across cultures, as illustrated by the Raven Progressive Matrices in 1938 and the Cattell Culture Fair Test in 1940. Such culture-reduced tests consisted of items intended to be equally familiar or unfamiliar to examinees of different backgrounds. (See Jensen, 1980, for a more complete discussion.)

The emphasis of the 1960s on equal opportunity in education and employment encouraged a systematic and sustained examination of test fairness. Cleary's seminal 1968 article on racial differences in scores on college entrance exams introduced an "equal regression lines" model as a standard for judging fairness in selection. Empirical studies using the Cleary model often found that it would sometimes

over-predict the performance of an ethnic minority group; that is, the group members would perform less well on the criterion than expected. The criterion performance of females, however, was often under-predicted, as the model suggested. Several competing definitions and models of unbiasedness appeared in the next few years, notably those of Darlington (1971), Thorndike (1971), Cole (1973), and Einhorn and Bass (1971). The debate was largely ended by Petersen and Novick (1976), who concluded that none of the models could be preferred solely on technical grounds and that value systems must inevitably enter into the choice of method.

Item Performance Methods

Many researchers gradually shifted from the summative evaluation of an entire test for fairness to the formative approach of studying particular items and item types, in the hope of building instruments without hidden inequities. (See Table 2.1 for a conceptual framework of approaches to bias.) Probably the system most commonly adopted by test developers was the formal institution of judgmental reviews of item and test content (Tittle, 1982). Test specifications, item writing and review, and final item selection are all be stages of test development at which the advice of reviewers and outside experts can be sought. Such judges check items for stereotyping, positive balance, cultural unfamiliarity, and congruence with curricula and opportunity to learn. Often standardized rating schemes and checklists are used both to train judges and to document results (e.g., Hunter & Slaughter, 1980). Although judgmental bias reviews do not correlate well with statistical ones (Schmeiser, 1985), the use of judgmental methods

TABLE 2.1

A CONCEPTUAL FRAMEWORK OF BIAS

Methodology	Measurement Level: Item	Measurement Level: Test
Qualitative	offensive language stereotypes face validity differential familiarity item format	specifications balance and representation composition of tryout sample test directions examiner effects reading level
Quantitative	difficulty discrimination distractor analysis differential item performance	predictive validity criterion adequacy standard-setting

Adapted from textual material by Schmeiser (1985).

is still important to insure procedural fairness.

Another branch of item bias research used empirical or statistical methods to examine items for differential performance. Early articles involved analytical methods for studying item difficulty-by-group interactions (Cardell & Coffman, 1964; Cleary & Hilton, 1968). The first method to gain widespread popularity was the delta-plot method (Angoff, 1972; Angoff & Ford, 1973), today called transformed item difficulty or TID. In this method, item p values are calculated for two groups, converted to normal deviates (usually deltas, $d = 4z + 13$), and plotted on a graph. The items falling at the greatest distance from the major axis of the scatterplot show the greatest group differences in relative difficulty. There are several variations on TID: testing the distribution in the differences for delta for normality (Echternacht, 1974); transforming p values to within-group standard scores and measuring their distance from a 45-degree major axis (Rudner, Getson, & Knight, 1980); calculating rank-order correlations of delta decrements, the difference in deltas between items (Jensen, 1980); and partialing out true score before calculating correlations (Stricker, 1982). The major limitation of TID is its confounding of item difficulty with item discrimination, especially when the groups under consideration have different ability levels, with the consequence that highly discriminating items are flagged erroneously (Angoff, 1982). Shepard, Camilli, and Williams (1985) modified the technique by regressing Angoff bias statistics on their point-biserials and then calculating residual delta indices.

Since 1970 more than a dozen other techniques for detecting differential item performance have been proposed. Some of them were

rejected on theoretical or empirical grounds; for example, the point-biserial item-test correlation procedure (Green & Draper, 1972) was shown to have artifactual problems relating to ability distributions (Hunter, 1975) and also performed poorly in several comparative studies discussed below. Other techniques, such as the Del statistic (Pennock-Roman, 1983), failed to get the attention of many researchers and simply passed from view without evaluation. Two methods, actually two families of methods, did attain acceptance and dominated research on item bias through the mid-1980s: chi-square and item response theory.

The chi-square method, originally presented by Scheuneman (1975, 1979), was later expanded by Camilli (1979) and Marascuilo and Slaughter (1981). It was the first procedure for detecting item bias that controlled for ability. Examinees are separated into several ability levels on the basis of their observed score on the total test or subtest. Within each ability level, the expected number of examinee responses to an item is compared to the actual number of responses for that group, and a goodness-of-fit statistic is tested. An item is considered unbiased when all persons of a given ability level have an equal probability of a correct item response regardless of group membership. Chi-square techniques, like other item bias techniques, assume that the ability being measured is homogenous and that the total test or subtest score is a reasonable measure of it. Chi-square techniques do not require normality or a constant direction of bias, although they can be affected by highly dissimilar ability distributions, by greatly unequal numbers of examinees per group, and by unreliability of the total test score (Scheuneman, 1976). The

chi-square methods are discussed at greater length in Chapter 3.

The chi-square procedures were developed independently of item response theory but may be viewed as approximations of it, using discrete intervals of observed ability instead of continuous curves of latent ability (Rudner, Getson, & Knight, 1980). The three-parameter model (hereafter IRT-3) uses difficulty, discrimination, and guessing parameters to describe the ability curve. The parameters and ability levels are found through an iterative maximum likelihood procedure requiring special computer programming. Biased items will have nonequivalent curves for different groups. The indices used to measure bias include several ways of computing the area between curves and a test for the equality of parameters across groups (evaluated by Shepard, Camilli, & Williams, 1984). IRT-3 requires sample sizes of 1000 or more per group.

The "pseudo-IRT" method (Linn & Harnisch, 1981) and one-parameter IRT can be used with smaller samples. The pseudo-IRT method uses three-parameter IRT on the total group of subjects to obtain estimated values for the probability of a correct answer, which are then compared to the actual values for each group. The one-parameter or Rasch model (hereafter IRT-1) permits only the difficulty parameter to vary. Bias in an item is shown by the area method, the difference in difficulty, or the mean square fit statistic (Durovic, 1975; Wright, Mead, and Draba, 1976). IRT-1 practitioners have also developed techniques for identifying individual persons who do not fit the model. (See Ironson, 1982b, for a more complete treatment of chi-square and item response theory methods.)

Comparative Studies

The decade from 1975 to 1985 produced at least two dozen comparative studies examining the relationships among bias detection methods and measuring their success in identifying differential item performance. Table 2.2 lists a selection of these studies.

The simulation and induced-bias studies, designed to measure the accuracy of methods in finding items known to be biased, for the most part come to the same conclusions as the empirical research studies assessing the concordance among methods. The IRT-3 methods are generally preferred by the studies that use it on the grounds of theory (because of the statistical independence of persons and items) and of psychometric behavior. The simulation studies have used IRT-3 procedures to generate their data and hence offer IRT methods an advantage, but the real-data research also supports the IRT methods. These techniques, however, are expensive, complex to implement and interpret, and demand sample sizes unrealistic in many testing situations. The pseudo-IRT approach is simpler but still requires extensive computer support. When IRT calibration is impractical, the full chi-square (X2) has been the method of choice (Subkoviak, Mack, Ironson, & Craig, 1984; Shepard, Camilli, & Williams, 1985). The TID method is not as useful, although using residualized deltas can bring its performance close to that of X2 (also Shepard, Camilli, & Williams, 1985). The IRT-1 methods, like the original TID, are adversely affected by variance in item discriminations and are not recommended (Shepard, Camilli, & Averill, 1981).

TABLE 2.2
COMPARATIVE STUDIES OF ITEM BIAS DETECTION TECHNIQUES

Author & Date	Test	Groups Compared	Smallest Group	Ability Difference	Methods Compared
Beck & Sklar (1978)	Metropolitan Achievement	national norm group & urban (minority)			TID, point-biserial, distractor analysis
Rudner & Convey (1978)	Stanford Achievement	"normal" and hearing impaired	1600	.6 s.d.	factor analysis TID, C2, IRT-3
Ironson & Subkoviak (1979)	National Longitudinal Study	White/Black high school seniors	1700		TID, point-biserial, C2, IRT-3
Merz & Grossen (1979)	simulation	----------	1000		TID, point-biserial, C2, IRT-3, IRT-1, factor analysis
Rudner, Getson, Knight (1980)	simulation	----------	1200	1 s.d.	TID, C2, IRT-1, IRT-3
Shepard, Camilli, & Averill (1981)	Lorge-Thorndike Intelligence	White/Black/Hispanic 4th - 6th graders	500	.7 s.d.	TID,point-biserial, C2, X2, IRT-3, IRT-1
Stricker (1981)	GRE Verbal	White/Black, M/F college students	300		TID, IRT-3, partial correlation
Burrill (1982)	Metropolitan Readiness Test	White/Black first graders			TID, point-biserial, C2, 4 other difficulty/ discrimination
Ironson, Homan, Willis, & Singer (1984)	math test w/ planted items	good & poor readers in 2nd & 4th grades	150		TID, X2, pseudo-IRT

TABLE 2.2 (continued)

Author & Date	Test	Groups Compared	Smallest Group	Ability Difference	Methods Compared
Shepard, Camilli, & Williams (1984)	High School & Beyond data	White/Black teenagers	3000	.9 s.d.	IRT-3 (7 types)
Subkoviak, Mack, Ironson, & Craig (1984)	College Qual. Test with planted items	White/Black college students	1000		TID, C2, X2, IRT-3
Raju & Normand (1985)	SRA Achievement	White/Black middle school students	1000	.75-1 s.d.	TID, point-biserial, C2, X2, IRT-1, IRT-3, regression
Shepard, Camilli, & Williams (1985)	simulation	----------	300	.8 s.d.	TID, X2, IRT-3, psuedo-IRT

Recent Developments

Three techniques, too recently proposed for inclusion in the comparative studies cited, all belong to the classical conditional probability paradigm. The standardization method (Dorans & Kulick, 1986) generates expected frequencies for each point on the observed score scale based on the performance of the base or reference group. The actual performance of the contrasted, or focal, group is then compared to the expected frequency, and the difference in p is standardized by a common weighting factor at each score level (unlike chi-square, in which each group is weighted by its own relative frequency). As with chi-square and IRT methods, both signed and unsigned summary bias indices can be generated. The standardization method, however, requires a sample size that can be even larger than IRT-3.

The Mantel-Haenszel method, as applied to item bias, compares the odds that the reference group at each score level will get an item correct to the odds of the focal group doing so. The odds ratio is weighted and transformed in various ways to yield statistics which measure the amount of differential item performance (Holland & Thayer, 1986). The Mantel-Haenszel statistic is relatively easy and inexpensive to calculate, and it is rapidly gaining acceptance. At present, however, it cannot accurately measure disordinal effects such as those seen when item characteristic curves cross.

The third new conditional probability method is log-linear analysis (Mellenbergh, 1982). The data for each item on a test can be displayed in a multidimensional table (ability level by group by response). The natural logarithm of the ratio of correct and incorrect responses for a

given ability and group is called its logit. The logit model for an unbiased item needs only parameters for an item constant and the ability level. A biased item will require the addition of group parameters, and possibly ability by group interactions, to obtain a model that fits the data well. This method is also discussed more fully in Chapter 3.

Selecting a Method

No empirical bias detection methodology now in use possesses all the desirable properties set forth by Ironson (1982a). Theoretically, an item bias statistic should have a known sampling distribution to allow significance testing and should be powerful, robust, and free from artifacts. Psychometrically, it should be reliable and have construct validity. Practically, it should be easy to calculate and interpret, have wide availability, cost relatively little, and be easily understood by test users. Further, the removal of items identified by the technique should not have drastic effects on the reliability and validity of the revised test.

As Table 2.2 shows, most of the comparative studies to date used norm-referenced tests or simulations in which the mean difficulty was near average and the score distribution did not depart wildly from normality. The studies did cover a variety of content areas and item formats. As for subjects, most research compared two sizeable groups, usually differing in ethnicity, and with moderate differences in mean ability (up to 1 s.d.). More studies are needed to test the limits of the detection methods, e.g. on highly skewed or bimodal tests, or if unidimensionality is violated, or with more groups, or smaller ability

differences. More work is also needed on the stability, reliability, and robustness of bias techniques, such as the studies by Hoover and Kolen (1984) and Harris and Hoover (1986), which suggested that existing techniques may be overly influenced by chance factors.

CHAPTER 3

METHODOLOGY

This exploratory study was concerned with two main issues in the area of methodological detection of differential item performance.

Question 1: How do two chi-square methods, X2 and C2, compare when applied to a minimum competency test?

Question 2: How do chi-square and log-linear methods of detecting differential item performance compare when applied to a minimum competency test?

It also considered three secondary substantive issues.

Question 3: Is there evidence of differential item performance by ethnic group on the MEAP Grade 4 reading and mathematics tests? If so, are there any interpretable patterns?

Question 4: Is there evidence of differential item performance by language group on the MEAP Grade 4 reading and mathematics tests? If so, are there any interpretable patterns?

Question 5: Is there evidence of differential item performance by sex on the MEAP Grade 4 reading and mathematics test? If so, are there any interpretable patterns?

This study differs from earlier research in comparing X2 and C2 on the kind of test for which they were designed and in taking advantage of the multivariate nature of log-linear analysis to fit multiple models and look for interactions among sex, ethnicity, and ability.

The Michigan Educational Assessment Program

Under the direction of the State Board of Education, the Michigan Educational Assessment Program (MEAP) carries out testing in grades 4, 7, and 10 to provide information on the status and progress of basic skills education in the public schools. Every student (except for the exclusions mentioned in the Subjects section) in the above grades is tested each fall on selected minimal performance objectives in reading and mathematics. (In addition, a random subsample of schools is used to assess achievement in other areas, e.g. science, music, and career education.)

The MEAP tests are criterion-referenced and objective-based instruments. Each objective is measured by three multiple-choice items; the student must answer at least two correctly to pass the objective. Students may pass or fail objectives but do not pass or fail the test as a whole. The score reports do place students into one of four achievement categories for each subject based on the percentage of total objectives attained, with three-quarters of the students falling into the highest category. In the current educational climate, some local school districts are moving towards using MEAP total test scores or achievement categories for decisions about grade promotion and high school graduation. This "high-stakes" extension of MEAP's impact heightens the importance of ensuring its equity and fairness.

The MEAP tests have been constructed according to professional standards. Involved in the process for the present tests were the technical staff of the Michigan Department of Education, the Michigan Reading Association, the Michigan Council of Teachers of Mathematics, and educators from local districts, who reviewed all objectives and

items for their quality and content validity. Detailed statistical analysis has been performed at the item and objective level annually (Phelps et al., 1980).

Because the total test score is not used for instructional purposes, the Michigan Department of Education has not conducted factor analysis or reported traditional test statistics at the overall test level (Roeber, personal communication, April 1988). As evidence of unidimensionality, the correlation coefficients of performance on each objective with performance on all objectives range from .45 to .70 for Grade 4 reading (median .63) and from .26 to .62 for Grade 4 mathematics (median .53) (Phelps et al., 1981). For this study the researcher calculated KR-21 to be .93 for Grade 4 reading and .94 for Grade 4 mathematics, using a preliminary sample of 4430 students and including all core and supplementary objectives. The tests, although not perfectly unidimensional, appear to have enough internal consistency so that total test score can be used to estimate ability.

The objectives and items underwent judgmental review for sex bias and stereotyping, using the Macmillan Guidelines (Macmillan, 1975), before regular administration of the test began (Phelps et al., 1980). The items have usually not been examined statistically for differential item performance between sexes, ethnic groups, or language groups. In fact, information on student characteristics (except for sex) is not routinely collected, presumably since such factors should not affect planning for basic skills instruction. MEAP staff did conduct a pilot study (Roeber, 1984) with six volunteer school districts and found differences in objective attainment among ethnic groups, with Black and Hispanic examinees receiving lower scores, especially in reading.

The 1984-85 Grade 4 MEAP test had 25 core objectives in reading and 28 core objectives in mathematics, as well as supplementary objectives not analyzed in this study. As each objective was measured by three items, there were 75 core reading items and 84 core mathematics items. The reading skill areas included vocabulary, comprehension, and study skills; the mathematics skill areas included numeration, whole numbers, fractions, measurement, and geometry. The mathematics items required very little reading ability. The objectives were intended to measure minimal skills and hence proved quite easy for most students, with statewide difficulties usually over .80 (Phelps et al., 1981). (See the State Summary Report in Appendix A for a complete list of objectives and the statewide percentage of examinees who passed each one.)

Subjects

As already mentioned, the MEAP test population consisted of all the fourth, seventh, and tenth graders enrolled in Michigan public schools when the test was given in late September of each year. Students absent during the scheduled testing were supposed to make it up, and those repeating a grade also repeated the test. At the tenth grade level the percentage of students participating has been much lower than expected in some high schools, especially in urban districts (Roeber, 1984). Broadly speaking, however, MEAP test results could be generalized to all Michigan students in the target grades.

Only two types of students could be excluded from MEAP testing. The first was students receiving more than 50% of their reading/English instruction in special education programs (e.g. mentally

impaired, emotionally impaired, learning disabled, or physically unable to take the test). The other category contained students from non-English-speaking countries who had been enrolled in U.S. schools for less than a year (these were usually from Southeast Asia or the Middle East). Schools had to report the total number of pupils excluded from MEAP testing but not the reason for doing so. The numbers actually reported in past years suggested that some schools may have excluded children that should have received testing.

The sample for this study used only fourth grade students in order to reduce the effect of within-school curricular differences, because variance between elementary classrooms presumably is less extreme than that between high school schedules. The restriction to fourth grade should also have improved the accuracy of the teachers' assessment of pupils' ethnic and language status, because elementary teachers would spend more time with each student. The study was further limited to the MEAP reading and mathematics testing, as other academic subjects were tested only on samples of the school population.

Because ethnic and linguistic groups were not evenly distributed throughout the state, selective sampling was preferred to random sampling in order to obtain a reasonable number of students from minority groups while keeping overall sample size to manageable proportions. The Hispanic group, less than 3% of the state total, was the most limiting factor in selection. The latest available school racial-ethnic reports were examined to identify the districts with the greatest numbers or highest percentages of minority students. Forty districts reported 750 or more minority students in grades K-12; of these, the 22 districts with at least 180 Hispanic students in grades

K-12 were selected for further study.

The school building records on computer file gave enrollment by grade and racial composition for each building within a district, enabling the researcher to find buildings likely to have at least five Hispanic students in fourth grade. These buildings were checked against the latest list of school closings and the MEAP special subject area test sites. The final sample chosen for study had 77 schools in 14 districts and contained about 4865 students in the 1982-83 school year, of whom 50% were White, 30% Black, and 15% Hispanic. (The method of racial/ethnic classification is discussed below.) All buildings selected had both Anglo and Hispanic students; the proportion of Black students ranged from none to a majority. Most buildings had students from all three ethnic groups.

All the districts chosen were sent a letter, cosigned by an official of the Michigan Department of Education and the researcher, explaining the study and asking them to participate. Every district agreed to be part of the study; four large ones, with about half the total sample of students, could not provide language proficiency data because of technical considerations such as pre-gridded answer sheets. The districts were mostly urban or urban fringe, though some were relatively small (5000 students K-12) or rural. Socioeconomic variability was presumably lower than the state average, especially for majority students, since rural and wealthy suburban districts were less likely to be included. This reduction in variability was furthered by using the school building as the sampling unit, since most of the schools were neighborhood-based. Likewise, the differences between ethnic groups in curriculum and opportunity to learn were diminished.

As required, clearance was sought and received from the University Committee on Research Involving Human Subjects. The study was given a Type 2 exemption applying to the use of educational tests in such a manner that subjects cannot be identified.

Because differential item performance may be quite small in terms of effect size, a sample size large enough to have appropriate statistical power was desirable. The sample chosen was estimated to have a power of over .95, so that the possibility of committing a Type II error was very low (Cohen, 1969). A large sample size would, of course, increase the probability of finding statistical significance even when the real impact of a difference was minimal.

In light of the minimum cell size needed for meaningful results and the practical constraints of the research situation, the study was not designed to analyze ethnicity effects for Native Americans and Asians or language-ethnicity interactions for most groups. The study was designed to analyze sex effects; ethnicity effects for Blacks, Hispanics, and Whites; and language effect for Hispanics.

Instruments

Because this study was largely methodological, the MEAP test described above could be considered as the object of study rather than as an instrument. The research instruments were the demographic survey and the methods chosen to detect differential item performance. This section discusses the reliability, validity, and objectivity of the demographic survey; the techniques section discusses the detection methods.

The demographic survey consisted of finding out the sex, ethnicity,

and language dominance of each student in the sample. Sex was determined by student self-report on the MEAP answer sheet; this information had been collected routinely for years without difficulty and met the three requirements above.

Schools had to report ethnic group for every enrolled student on their "Fourth Friday" forms. They used the standard Federal classifications set forth in an OMB directive (Office of Management and Budget, 1979), which defined five groups summarized as follows:

1. American Indian or Alaskan Native
2. Black, not Hispanic
3. Asian or Pacific Islander
4. Hispanic
5. White, not Hispanic

"Hispanic" was an ethnic label, not a racial one; for example, Dominicans (Black), Colombians (White), and Mexicans (often Indian) were all classed as Hispanics. Filipinos were included with Asians.

The form provided to Michigan schools directed them to include a student "in the group to which he or she appears to belong (or) identifies with." Problems with reliability, validity, and objectivity could arise in several ways. The judge might make an erroneous decision; different judges might classify the same student differently; and no procedure is offered for resolving multiple group membership. Nonetheless, the Federal categories seemed the best choice for this study. The information was already required by the state and federal governments and was collected about the same time of year as the MEAP testing. The pilot project previously mentioned, the Hispanic Coding Study (Roeber, 1984), used these categories and had a coding

error rate of less than 2%. Also, for political and practical reasons it was preferable to work within the existing system.

The most difficult aspect of the demographic survey was the collection of information on language status. Although bilingual and migrant education programs made some determination of individual students' language abilities, there was apparently no standard statewide procedure for doing so, nor did the regular classroom teachers have a standard classification system. This information was especially important for interpreting the MEAP test results for Hispanic children, since few Hispanics met the formal criteria for exclusion from MEAP testing; students who had attended school in the United States for twelve months or more were supposed to be tested regardless of language ability.

In this study, language information was collected by asking the teacher or administrator to determine each student's best language at school. The staff member used a two-way table to find the single code that represented a student's ethnicity and language dominance and then gridded that code in the research block on the student's MEAP answer sheet. For example, a "14" represented an English-dominant Hispanic, and a "19" an other-dominant Hispanic. The language proficiency category was dichotomous: a) English-monolingual or English-dominant; b) monolingual or dominant in another language. The other language was not specified. This approach minimized the demand on the coder to make detailed judgments.

Procedure

After the school districts described in the "Subjects" section had agreed to participate, they received a mailing explaining how to code and report each student's ethnic and linguistic group. The coder (usually a staff member of the central district office) gridded in the two-digit ethnicity-language code. Districts that could not supply language information used a specified single-digit number for ethnicity. In most instances the gridding was done before MEAP tests were given. In two cases the researcher did the gridding at the district office after the MEAP tests had been given. Students recorded their own sex and birthdate at the time of testing. The school districts also completed a form for the researcher indicating the source of their information and explaining any difficulties.

The actual MEAP testing proceeded in the customary manner. When each district had completed testing, the answer sheets were sent to the contracted scoring service according to standard procedure. The contractor then prepared a special tape for the schools in the research sample, including all the data except student name.

The research data was first analyzed for demographic variables with SPSS. Records were received for 4430 students. Students with any missing demographic data (8%), mostly the result of district error, were excluded from the final sample, as were Asian and American Indian students (3%) because of small numbers. The median age for all students was 9 years 7 months; a large number of examinees (6%) were older than 10 years 10 months, that is, more than one year over age for grade. Hispanic males were more likely to be older and Black males younger. Because this characteristic was not randomly distributed

across sexes or ethnic groups ($p < .001$), these students were also dropped from the sample in order to avoid the confounding of sex or ethnicity with delayed entrance or retention. In total, about 17% of the original sample was excluded from the final analysis.

The final sample consisted of 3695 students, of whom 52% were girls and 52% boys. The percentage of students in each ethnic group remained nearly the same as in the original sample: 54% White, 30% Black, and 17% Hispanic. (Table 3.1 gives the exact cell counts.) There was no significant association between ethnicity and sex; the percentage of girls was 52% for Whites, 53% for Blacks, and 51% for Hispanics. The sample represented 3.5% of all MEAP Grade 4 examinees. It should be noted that, according to 1980 federal census figures for Michigan (Census Bureau, 1983), Hispanics accounted for 2.5% of this age group. The final sample in this study thus was estimated to contain approximately 24% of the estimated 2600 Hispanic fourth graders statewide, a significant proportion.

Of the 292 Hispanics for whom language information was available, 79% were judged to be English-dominant. This finding was supported by the 1980 census results (Census Bureau, 1983) showing that approximately 82% of all Hispanic residents of Michigan were born in the United States and that almost half of the families did not report currently speaking Spanish at home. Because there were only 61 other-dominant (probably Spanish-speaking) Hispanic students, the Hispanic group was not subdivided for language analyses, and Question 4 about the effect of language group on performance could not be addressed.

TABLE 3.1
COMPOSITION OF FINAL RESEARCH SAMPLE

	Black	Hispanic	White	Total
Male	514	308	952	1774 (48%)
Female	576	321	1024	1921 (52%)
Total	1090 (30%)	629 (17%)	1976 (54%)	3695 (100%)

Techniques

The item bias techniques chosen should be appropriate for examining the relationship of the independent nominal categories of group membership (sex, ethnicity) to the dependent variable of a dichotomous item response (item correct or incorrect), taking ability into account. If ability can be satisfactorily measured on an interval scale, then large-sample approaches such as IRT-3 or standardization are possible. This study treats ability as a categorical variable by dividing the continuous variable of total test score into five levels, an approach which permits smaller sample sizes.

Because of the nature of the test and the examinee population, it is unwise to assume a normal distribution of ability or homogeneity of variance. The MEAP reading and mathematics score distributions are quite negatively skewed, as indeed would normally be expected for criterion-referenced tests administered following instruction in the content to be tested. The ability distribution for a group may thus be truncated, and the size of any true difference in ability between groups may be distorted (Loyd, 1986). Mastery or criterion tests may

cause difficulty for many item bias methods (Scheuneman, 1980).

An appropriate analytic technique for this kind of situation is multidimensional contingency table analysis with chi-square test statistics (Andrews et al., 1981). The three item bias techniques chosen for this research study fall into this category. They will be discussed here in order of the date they were first suggested for use in item bias studies, which also happens to be in order of increasing statistical complexity.

The Scheuneman C2

The first step in calculating C2 or X2 is to establish the ability groups. Three to five intervals provide the best performance. There is not an algorithm for setting the ability levels; they may be set on the basis of width of score interval, number of people, or smallest cell frequency. (As Ironson (1982b) points out, the arbitrary nature of the ability levels is a disadvantage of all the contingency-table methods; treating ability as a categorical variable inevitably results in loss of information. These methods, however, may be the best solutions to handle small samples or for situations in which complex calculations are not feasible.) Scheuneman's original use of C2 set the ability level by dividing the distribution of correct responses for the smaller group into fourths or fifths (Scheuneman, 1976). In this study, ability levels were determined separately for reading and mathematics by dividing the total group of examinees into five approximately equal groups based on total test score for that subject. The same intervals were used for each item within a subject area instead of being allowed to vary across items; furthermore, these

intervals were also used for the X2 and log-linear calculations to standardize comparisons.

The second step is to calculate the chi-square value and test it for statistical significance. The Scheuneman C2 (so labeled because it does not have a true chi-square distribution) is computed using only correct responses. It can be represented as follows:

$$\text{if } m = 1: \quad C2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(Eij - Fij)^2}{Eij} \qquad \text{with } df = (I-1)(J-1)$$

where E = expected frequency of response
F = observed frequency of response
i = ability group
j = status group
m = item response (1 = correct)

The Full X2

The full X2 differs from C2 because it includes incorrect responses (with a corresponding loss of degrees of freedom). It can be represented as follows:

$$\text{for any } m: \quad X2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(Eij - Fij)^2}{Eij} \qquad \text{with } df = I(J-1)$$

where E = expected frequency of response
F = observed frequency of response
i = ability group
j = status group
m = item response

Both the C2 and the X2 techniques offer the advantage of being able to make several comparisons (e.g., among three or more ethnic groups) simultaneously, thereby reducing the amount of work involved and the risk of spurious results associated with conducting a large number of

significance tests. The techniques are also relatively simple to compute and understand.

The C2 has been criticized on theoretical grounds by Marascuilo and Slaughter (1981) and by Baker (1981a) because its exclusion of incorrect responses produces an unknown distribution which may not approximate the X2 distribution, especially if the group sizes are quite different or the cell frequencies are all very large. Marascuilo and Slaughter also pointed out that the C2 is an omnibus test of differences between expected and observed values (are all the differences equal to zero?). This is less efficient than the orthogonal pairwise comparisons of the full X2 (is a given difference equal to zero?).

Although C2 and X2 are highly correlated, X2 is usually favored in comparative studies. For a very easy item, however, such as those on a minimum competency test, the X2 may be inflated by the small number of high-ability examinees expected to miss the item. The C2 is less likely to identify an easy item as biased. In fact, it was developed for application to a very easy test (Scheueneman, 1980). The C2 also does not require as large a sample size; it has been used with only 150 subjects in the smaller group and with scales as short as 10 items. (For a thorough comparison see Ironson, 1982.)

The application of either the C2 or the X2 to an item results in a chi-square statistic to be tested. An item may be labelled as "biased" (differentially performing) in several ways:

1. The chi-square statistic exceeds a predetermined significance level such as .01, .05, or .10. Scheueneman (1976) argues for using a more liberal level, even .20 or .30, in order to identify

trends and patterns more easily.

2. The item falls into a predetermined position or category (the most biased, the top 5%, etc.).

3. The chi-square statistic exceeds the one reaching significance in a random comparison (e.g. White-White) or a pseudo-group comparison (e.g. random males vs. males with the same ability as females). This approach reduces the effect of random noise in identifying items as biased and compensates for the lack of a known distribution.

A special FORTRAN computer program was written for this study to calculate both the C2 and X2 statistics in a single run. The program also tested each cell size to insure that the expected frequency was adequate (n = 1 or more).

Log-linear analysis

This section is based on the writings of Baker (1981b) and Kennedy (1983). For a more theoretical discussion of this complex area, see Feinberg (1977).

The chi-square methods just described permit analysis of a two-dimensional contingency table (status group by item response) for each level of ability. In order to examine designs of greater dimensionality, such as two independent types of status groups, the more sophisticated log-linear method is needed. An early article suggesting this method (Mellenburgh, 1982) reformatted Scheuneman's chi-square in terms of a log-linear model. Log-linear analysis resembles analysis of variance in many ways. (Unlike ANOVA, log-linear models deal in the frequencies of a variable, not its values, and they

do not contain an error term.) It produces a linear model, rather than a multiplicative one, by taking the natural logarithm of the expected frequencies:

$$\ln Fij = \ln \left(\frac{Fi \; Fj}{N} \right) =$$

$$= - \ln N + \ln Fi. + \ln F.j =$$

$$= U + Ui + Uj$$

where U = constant frequency
Ui = main effect for i
Uj = main effect for j

The fully saturated model, U + Ui + Uj + Uij , explains ln Fij completely. In log-linear analysis it is possible to set up models representing the variables of interest, then fit each model of interest and obtain a residual goodness-of-fit measure to determine the overall agreement of the model with the observed data. The preferred goodness-of-fit measure is a maximum likelihood, G2, but Pearson X2 can also be used. (The contribution of an individual term in the model to the goodness of fit can be obtained by examining the difference in fit, called the component, between a model with the term and one without it.) Examination of residuals to find the best model proceeds from the most complex model down, in order to simplify interpretation by tending to higher-order associations first (Baker, 1981b).

The approach used in this study is technically a logit-linear one. The logit-linear technique applies when the observations are sampled from multiple populations and when there is a distinction between explanatory (independent) variables and response (dependent) variables. In other words, logit-linear analysis is used for asymmetrical designs that identify differences between groups with

respect to their responses (Kennedy, 1983). Logit-linear analysis uses only a subset of the models possible in the corresponding log-linear situation, because the response variable must always be included in the model.

In the detection of item bias, if a logit-linear model including only the constant U and the ability term fits the data, then there is no bias present. If a term must be added for status group membership, there is uniform bias. If the ability-status group interaction term is required, there is non-uniform bias -- bias against low-scoring members of one group and high-scoring members of another group (Ironson, 1982b).

Dutch researchers have applied the log-linear procedure to simulated data (van der Flier et al., 1984) and to real data with experimentally induced bias (Kok, Mellenbergh, & van der Flier, 1985). They tested the fit only of the model for an unbiased item (see Model 2 below). In both cases they used an iterative method, excluding the items with the highest G2 values until all remaining items had nonsignificant G2 statistics.

In the United States, one study (Alderman & Holland, 1981) computed G2 to check for interactions of ability and language groups as part of a chi-square study. Loyd (1984) investigated stability of the index across samples. In another study (Loyd, 1985), the log-linear method was compared with two IRT-1 methods on a minimum competency test (average item difficulty = .83). Loyd, like the Dutch researchers, tested the fit of only the unbiased model. She used multiple samples of varying size to check stability of the indices and of item classifications.

This study used the MULTIQUAL program (Bock, 1973) to compute both G2 and Pearson X2 statistics. The program must be run separately for each item. Because the appropriate technique was logit-linear, there were eight possible models to fit:

Model	Terms Included	Interpretation
1	U	constant only
2	U,I	constant and ability
3	U,I,J	constant, ability, and one status group
4	U,I,J,K	constant, ability, and both status groups
5	U,I,J,K,IJ	ability-status interaction
6	U,I,J,K,IJ,IK	two ability-status interactions
7	U,I,J,K,IJ,IK,JK	above plus status-status interactions
8	U,I,J,K,IJ,IK,JK,IJK	three-way interactions (saturated model)

The choice of model depends on the researcher's judgment in cases where the significant residual and significant component do not agree on a single model (Kennedy, 1983). To standardize the analysis in this study (so that replication would be possible), the model chosen was always the most complex one indicated by a significant residual.

Assumptions

The methods chosen in this study for detecting differential item performance assume that the total test is valid and relatively homogenous and that the total test score can be used to indicate ability. They do not make any assumption about the distribution of observed scores, including normality. In fact, these methods are well

suited to situations where the observed proportions are extreme (Scheuneman, 1976; Kennedy, 1983), such as a minimum competency test like the MEAP.

Although it is reasonable to assume that the total test score is a valid measure of ability, there are some problems (Ironson, 1982b). For one thing, the total score may not be free from measurement error. More seriously, the presence of biased items may contaminate the total score so that it underestimates the ability of the group that the items are biased against. All the items will then be more biased than the index shows, and this constant bias will become part of the scale. All item bias techniques based on classical test theory share this weakness. The techniques, however, can still indicate relative bias.

The study also does not make any theoretical assumptions about the causes of differential item performance, whether they are environmental, educational, or testing-induced. It is concerned with finding the most suitable ways of identifying whatever differences may exist. Furthermore, the fact that an item exhibits a statistical discrepancy does not necessarily mean that the item should be automatically discarded. Content validity and other qualities should also enter into the decision to accept or reject an item.

CHAPTER 4

ANALYSIS AND RESULTS

The preceding chapters described the problem of differential item performance, its background and context, and the specific procedures for conducting this study. This chapter presents a description of the results of the study.

The chapter is organized into three sections. First, the general performance of the sample of students is presented. Second, the results of both chi-square techniques and of the log-linear technique for detecting differential item performance are provided, and their relationships are explored. Finally, the content of the MEAP test is examined for patterns detected by the item bias techniques.

Performance of the Sample

The 3695 students in the final sample achieved slightly lower scores on the MEAP reading test and marginally lower scores on the mathematics test than did the statewide population (see Table 4.1). The differences between the sample and the statewide groups were greatest on the most difficult items.

Despite the somewhat lower scores, the sample distribution still was clearly that of a mastery test. The mean number of items correct was 75% for reading and 83% for math, and the modal number of items correct was 89% and 95% respectively. The modal number of objectives passed was the highest possible (25 of 25 reading objectives, and 28 of 28 mathematics ones).

TABLE 4.1
MEAN ITEM DIFFICULTY FOR STATEWIDE AND SAMPLE GROUPS

	Statewide n = 104,914		Sample n = 3,695	
	Reading	Math	Reading	Math
Mean	.79	.85	.74	.85
Range	.51-.93	.68-.99	.43-.93	.66-.99

TABLE 4.2
ANALYSIS OF VARIANCE OF TOTAL SCORE

	Source	df	Sum of squares	Mean squares	F ratio	prob.
Reading	Ethnicity	2	36633	18317	91.93	.0000
	Residual	3692	735615	199		
	Total	3694	772248			
Reading	Sex	1	5103	5103	24.56	.0000
	Residual	3693	767145	208		
	Total	3694	772248			
Math	Ethnicity	2	14357	7269	65.08	.0000
	Residual	3692	412376	112		
	Total	3694	426913			
Math	Sex	1	84	84	0.72	n.s.
	Residual	3693	426829	116		
	Total	3694	426913			

The next issue to consider was the presence of overall differences in ability within the sample itself between ethnic or sex groups. (Language dominance, as explained in Chapter 3, could not be analyzed.) Analysis of variance showed a significant relationship between ethnicity and reading score, sex and reading score, and ethnicity and mathematics score (see Table 4.2). The relationship between sex and mathematics score was not significant. In all cases, Whites and females performed better than minorities and males.

Identification of Differentially Performing Items

The statistics mentioned above indicated that there was a strong effect of sex or ethnicity on overall test score. The next step was to determine whether any part of the effect could be attributed to differential item performance, holding total ability constant.

All three methods used in this study for detecting differential item performance require dividing the sample into several levels of ability. As mentioned in Chapter 3, ability levels were determined separately for reading and mathematics by dividing the total group of examinees into five approximately equal groups on the basis of total test score in each subject, with the lowest level designated as 1 (see Table 4.3). The overall test score was employed rather than only the score on the core objectives.

Differences in group means within ability levels can produce regression artifacts that cause the appearance of bias (Shepard, Camilli, & Averill, 1981). The distributions of total test score for each group within an interval were checked at five points within ability level 1 and at four points within level 5. The only serious

discrepancy occurred at the lower end of level 1 for mathematics, where Hispanic examinees were likely to score higher than the White and Black examinees. As the Hispanic group was the smallest, it was not desirable to achieve a match by eliminating subjects.

TABLE 4.3
SCORE RANGE FOR ABILITY LEVELS

	Reading		Mathematics	
Level	No. of Items Correct	Percent of Sample	No. of Items Correct	Percent of Sample
1	0-49	20	0-76	20
2	50-61	20	77-87	19
3	62-68	21	88-94	20
4	69-73	20	95-101	22
5	74-86	20	102-108	19

In small samples, the C2 and X2 cannot always use the same score intervals because, for instance, there may not be enough incorrect responses by high-ability examinees. This study did employ the same ability levels for each item and for all three methods. Differences in total score distribution can inflate the chi-square values; when identical intervals are used for each item, such inflation will be systematic, allowing the derived chi-square values to be used as a relative index of bias (Rudner & Convey, 1978).

The ability levels were entered into the FORTRAN program for calculating C2 and X2 and into the MULTIQUAL program for calculating log-linear models. Table 4.4 provides a sample of the output of each program; item 2 is unbiased, while item 38 shows differential performance for ethnicity by all three methods. The percentage table indicates the proportion of examinees in each cell passing the item.

Next the Scheuneman C2 and full X2 statistics are shown. Strictly speaking, the lower values of the C2 "cannot properly be compensated for by adjustments in degrees of freedom" (Shepard, Camilli, & Averill, 1981, p. 338), and their true distribution is not known. Nonetheless, the significance testing approach may give an overall indication of the amount of bias present.

The log-linear table requires a two-step interpretation; first, one eliminates every model with a significant residual likelihood (such models do not adequately fit the data). Secondly, one retains only the models contributing a significant component likelihood. Ideally, as in the two items shown, the steps converge on a single model. Frequently, however, several models are acceptable, and the researcher must judge which one to accept (Kennedy, 1983). In order to standardize the decision process for this study, for each item the most complex model acceptable by the residual likelihood approach was chosen, similar to the ANOVA procedure.

Next, the chi-square output showing percent correct at each ability level was inspected to determine whether any bias detected was uniform (consistently favoring one group) or non-uniform (favoring the less able members of one group and the more able members of another). Because three ethnic groups were being compared simultaneously, it was possible for an item to clearly favor one group without consistently ranking another group lowest, or vice versa. If the pattern was uninterpretable, for example, if females were favored at ability levels 1 and 4 only, the item was not counted as biased in subsequent analyses. In the log-linear analysis, uniform bias appeared as a main effect for ethnicity or sex, and non-uniform bias appeared as an

TABLE 4.4
SAMPLE RESULTS

Item 2	Item 38
Objective I-A	Objective I-E
The children were <u>unsure</u> of the answers.	Terry had a <u>tame</u> raccoon.
The best meaning for the prefix <u>un-</u> is	The opposite of <u>tame</u> is
A. by. B. not. C. away. D. into.	A. wild. B. gentle. C. trained. D. unfriendly.
Percent correct = 88%	Percent correct = 63%

Percent correct at each ability level

	1	2	3	4	5	1	2	3	4	5
White	66	86	93	96	98	42	51	70	84	92
Black	66	89	93	98	98	34	46	53	67	86
Hisp.	66	88	93	95	99	37	48	59	69	85
Female	65	87	92	96	98	35	47	61	76	88
Male	67	88	93	96	98	40	51	65	80	93

(continued)

TABLE 4.4 (CONTINUED)

	Item 2			Item 38	
Chi-square analyses					
Ethnic	C2 = 0.17	$p > .25$	df = 8	C2 = 16.37*	$p < .05$
	X2 = 3.03	$p > .25$	df = 10	X2 = 56.39*	$p < .001$
Sex	C2 = 0.15	$p > .25$	df = 4	C2 = 3.35	$p > .25$
	X2 = 0.83	$p > .25$	df = 5	X2 = 10.85	$p < .10$

Log-linear analyses

Model	df	residual G2	component G2	residual G2	component G2
item	29	449.59*	-----	677.05*	-----
A	25	19.32	430.22*	73.21*	603.84*
E	23	18.82	0.50	30.66	42.55*
S	22	18.31	0.51	22.56	8.10
AE	14	15.48	2.83	9.81	12.75
AS	10	15.22	0.26	7.51	2.30
ES	8	12.10	3.12	6.77	0.74
AES	0	.00	12.10	.00	6.77
Model accepted =		A		E	

A=Ability E=Ethnicity S=Sex

Note: Each log-linear model incorporates the ones above it. For instance, AE includes effects for the item constant, A, E, and S, as well as the AE interaction.

*significant at $p < .05$

interaction of ability with ethnicity or sex. Occasionally only the saturated model (three-way interaction) would fit the data; these items were still treated as biased, even though their interpretation is difficult. The item data appear in Appendix B.

Table 4.5 summarizes the results of the bias analyses. All the methods agreed that the reading test had a higher percentage of biased items than did the mathematics test. The methods also showed good agreement in the number and pattern of items identified as biased. The Scheuneman C2, the only method of the three that does not include incorrect responses in its calculation, identified fewer items as biased than the X2 and G2. Every item that was flagged by C2 had a corresponding X2 significant at $p < .001$ and a significant G2 as well; the C2 can be taken as an indicator of the most problematic items. There was a significant correlation between item bias and item difficulty; the easier items were significantly less likely to be biased by C2 ($r = .48$ to $.69$) and slightly less by X2 and log-linear indices ($r = .06$ to $.33$).

A more detailed analysis of the results shows which ethnic groups and sex were favored or disfavored by those items identified by both C2 and X2. The strongest effect occurred on the reading test, on which, according to X2, 11 items favored Whites and 3 more disfavored Blacks, while only 2 disfavored Whites. The results for sex on the reading test were mixed; 7 items favored females, 8 favored males, and one displayed non-uniform bias. The mathematics results for both ethnicity and sex showed a similar mixed pattern. For ethnicity, 4 items favored Whites, 5 favored Blacks, 4 favored Hispanics, 3 more disfavored Whites, and 1 more disfavored Blacks. For sex, 11 items favored

TABLE 4.5
NUMBER OF ITEMS IDENTIFIED AS BIASED

By Individual Method

Reading (75 items)

Type of bias	C2	X2	G2
None	68	43	47
Ethnic only	1	13	8
Sex only	1	10	1
Ethnic and sex	4	9	19

Mathematics (84 items)

Type of bias	C2	X2	G2
None	82	44	43
Ethnic only	0	14	13
Sex only	2	17	10
Ethnic and sex	0	9	18

By Combination of Methods

Reading (75 items)

Type of bias	C2/X2	C2/G2	X2/G2	All
None	43	47	37	35
Ethnic only	1	1	7	1
Sex only	2	1	1	1
Ethnic and sex	4	4	8	4
Percent agreement	67%	71%	71%	57%

Mathematics (84 items)

Type of bias	C2/X2	C2/G2	X2/G2	All
None	44	43	35	35
Ethnic only	0	0	9	0
Sex only	1	2	7	1
Ethnic and sex	0	0	6	0
Percent agreement	54%	54%	68%	43%

Note: Items identified if significant at $p < .05$.

females, 12 favored males, and one showed an ability-sex interaction.

Comparison of Methods

Table 4.6 presents the intercorrelations among the methods. C2 and X2 appear twice in each table because they must be run separately for ethnicity and sex. Spearman rank-order correlations are preferred (Shepard, Camilli, & Williams, 1984) because the Pearson product-moment correlation can be inflated or distorted by very extreme items. Both sets of correlations were lower in the mathematics test, in which overall bias was less; this result is expected because bias detection methods will not necessarily agree on the ranking of unbiased items. Also, all the correlations may have been reduced because the indices were unsigned; that is, items biased against any group fell in the same tail of the distribution (Shepard, Camilli, & Averill, 1981).

The correlations of methods between sex and ethnicity were, as one would expect, much lower than the correlations within each type. The Spearman rhos for C2 and X2 between type ranged from -.025 to .490, and within type ranged from .584 to .853. Because some items did have both ethnic and sex effects, a positive correlation would be expected even between groups. The correlations of C2 and X2 within group were quite reasonable, at .816, .853, .584, and .838. (The figure of .584 should be interpreted in view of the fact that C2 did not identify any mathematics items as having ethnic bias.)

The correlations between the log-linear technique and both of the chi-square approaches were significant, but not as high: they ranged from .487 to .758. As mentioned above, the correlations between X2 and G2 tended to be higher than those between C2 and G2. Both sets of

TABLE 4.6
CORRELATIONS BETWEEN THREE ITEM BIAS DETECTION METHODS

Reading

Pearson Correlation Coefficients

	C2 (E)	X2 (E)	C2 (S)	X2 (S)	G2
C2 (E)	----	.887	.583	.397	.831
X2 (E)		----	.452	.394	.900
C2 (S)			----	.837	.676
X2 (S)				----	.702

Spearman Correlation Coefficients

	C2 (E)	X2 (E)	C2 (S)	X2 (S)	G2
C2 (E)	----	.816	.490	.335	.697
X2 (E)		----	.306	.276	.758
C2 (S)			----	.853	.599
X2 (S)				----	.639

Mathematics

Pearson Correlation Coefficients

	C2 (E)	X2 (E)	C2 (S)	X2 (S)	G2
C2 (E)	----	.618	.359	.259	.589
X2 (E)		----	.031	-.005	.587
C2 (S)			----	.890	.618
X2 (S)				----	.659

Spearman Correlation Coefficients

	C2 (E)	X2 (E)	C2 (S)	X2 (S)	G2
C2 (E)	----	.584	.452	.239	.582
X2 (E)		----	-.010	-.025	.574
C2 (S)			----	.838	.487
X2 (S)				----	.539

C2 = Scheuneman chi-square
X2 = full chi-square
G2 = log-linear

E = Ethnic
S = Sex

correlations were probably affected by the fact that G2 could measure ethnicity and sex effects at the same time. It appears that the log-linear approach detected the same type of differential performance as the chi-square techniques, but it could not be considered interchangeable with them.

Patterns of Differential Item Performance

The results of the MEAP testing for individual students were generally interpreted at the level of objectives rather than items. Thus the presence of bias in items was of most concern when it affected two or three items for the same objective and thereby increased the probability of failure on that objective for a given group.

In this study, the distribution of differentially performing items was not random across objectives or content. Many objectives had no flagged items, while all three items of others were identified. Table 4.7 shows the objectives with at least two items identified by both X2 and G2. This table represents about one-fourth of the objectives but one-third to one-half of the identified items.

As was noted at the item level, the only strong pattern seemed to be that Whites were favored on reading content. This finding could perhaps be partially attributed to the relationship of reading skill to the nature of the home environment. The math items, more closely linked to direct instruction, in general did not favor any sex or ethnic group.

There was no clear reason why certain objectives appeared biased and others did not; for example, the items in Obj. I-A (prefixes) and Obj.

TABLE 4.7
OBJECTIVES WITH TWO OR MORE ITEMS IDENTIFIED AS BIASED

Obj	Bias	Favors	Content
		Reading (25 objectives total)	
I B	E,S	Whites; mixed sex	suffixes
I D	E,S	non-Blacks; mixed sex	synonyms
I E	E	Whites	antonyms
II G	E	mixed ethnic	likeness/difference
III B	S	mixed sex	cause/effect
III D	E	Whites	details
		Mathematics (28 objectives total)	
10-7	E	non-Blacks	fewest number
16-4	S	males	2-digit expansion
24-2	S	females	2-digit add, regroup
30-2	S	females	2-digit subtract
35-6	E,S	Blacks; males	multiplication
79-4/6	S	mixed ethnic; females	congruency

31-1 (subtraction with regrouping) were not flagged for bias, although Obj. I-B (suffixes) and Obj. 30-2 (subtraction without regrouping) each had all three items flagged for the same group by at least two methods.

Summary of Research Questions

Question 1: How well do the chi-square methods, X2 and C2, agree in measuring differential item performance on a minimum competency test?

The methods correlated quite highly, although the X2 identified many more items as biased. These additional items had difficulty values of .45 to .95 and hence were probably not overly affected by the statistical artifact for very easy items described earlier.

Question 2: How well do chi-square and log-linear methods for detecting differential item performance agree?

The two types of methods showed moderately high agreement. The log-linear method, however, differed enough that it cannot be considered a simple substitute for X2.

Questions 3, 4, and 5: Is there evidence of differential item performance by ethnicity, language, or sex on the MEAP Grade 4 reading and mathematics tests? If so, are there any interpretable patterns?

Language could not be studied because of the small number of other-dominant students. Nearly half the items were identified for ethnicity and sex by at least one of the three bias detection techniques, and around a third were identified by both X2 and G2. These numbers are much greater than chance and do provide evidence of differential item performance.

Although the flagged items tended to cluster into objectives, there were few clear patterns at either the item or objective level. An effect favoring Whites on the reading test was the strongest finding.

CHAPTER 5

CONCLUSIONS AND RECOMMENDATIONS

Efforts to maintain both equity and excellence in testing are prominent in our society. The problem of detecting differential item performance by ethnicity, sex, or other demographic characteristic continues to be important to test developers, test takers, and test users. Comparative studies suggest that three-parameter item response theory techniques are the most effective way to identify items that function differently for examinees of the same ability but of different backgrounds. There is still a need, however, for methods that can be used for small samples, in tryout testing, and in other circumstances where IRT is not applicable. The psychometric properties of a minimum competency test, such as many easy items and skewed score distributions, also place special demands on the method chosen.

One suitable family of techniques in this situation is contingency-table analysis, such as chi-square and log-linear measures. These techniques are nonparametric and thus appropriate for nonnormal distibutions and categorical variables. They have the added advantage of permitting more than two levels of a group to be compared at a time. Log-linear analysis also permits the simultaneous consideration of several independent variables and their interactions. It can show which main effect or interaction is contributing the most to item performance.

This study was conducted to study the relationships among three contingency-table methods, namely, two chi-square methods -- the Scheuneman C2 and the full X2 -- and the log-linear method. The

results of the Fall 1984 Grade 4 Michigan Educational Assessment Program reading and mathematics tests were examined for a selective sample of approximately 3700 students representing both sexes and three ethnic groups (Whites, Blacks, and Hispanics). The groups differed in overall performance on the tests; the study was intended to examine which of the small sample methods would best show whether individual items were differentiating among groups. A secondary goal was to study those items that were identified by one or more methods to see whether any particular content was associated with their anomalous behavior.

Conclusions

The C2 method, as expected, identified many fewer items than the X2. Indeed, on the mathematics test the C2 detected only 2 items biased for ethnic group, a chance level. What was unexpected was that the C2 statistic was more highly correlated with item difficulty than was the X2: the C2 was designed to handle very easy items and was supposed to be more resistant to the effects of random errors by high-ability examinees. All items that the C2 did identify were also selected by both the other methods. This congruence suggested that the C2 could serve as a "worst case" indicator of problem items.

The log-linear method was shown to correlate moderately well with the X2. Since the X2 had been recommended as the best method when three-parameter item response techniques are inappropriate, the high concordance of the log-linear with the X2 indicated that it too might be a reasonable choice in similar circumstances. The G2 statistic also offered a known distribution for significance testing. The log-linear method was quite efficient in that several types of groups and several

classes within each type may be scrutinized in one procedure. This ability to explore models of theoretical interest might mean that the log-linear method was worth using even when sample size or other factors made the three-parameter item response theory techniques possible.

Despite these advantages, the log-linear model did not seem to be a good candidate to use for routine screening of items in practical terms. The MULTIQUAL program had to be run separately for each item, and the selection of the best-fitting model for a complex design could require user judgment. Of more intrinsic concern was the fact that the method seemed oversensitive, that is, it found differential item performance effects to be statistically significant when the practical significance was minimal. For example, even the easiest item, with a difficulty of .99 or 1.00, still tested significant for ability at the .05 level.

As for substantive findings about the relationship of differential item performance to item content, this study replicated the commonly found pattern of mixed results that balance each other out. Except for the preponderance of reading items that strongly favored Whites, there were few interpretable results. Identified items did tend to cluster within objectives, but the meaning of that occurrence was unclear because apparently similar objectives did not demonstrate the same effects. Because the MEAP test is supposed to be a minimum competency one, measuring concepts heavily emphasized in the public school curriculum, the lack of substantive content factors should be reassuring in terms of the test's functioning.

Recommendations

As always, additional research on the topic of this study would be desirable. Because this study used real data, there was no independent criterion of item bias. A simulation study with generated bias or one using planted items designed to be biased would yield a more refined indicator of the accuracy of the log-linear technique. In particular, a simulation study using three-parameter item response theory to model item difficulty, discrimination, and guessing would enhance understanding of how item characteristics can affect the accuracy of all three contingency-table analysis methods.

It would also be interesting to compare the results obtained using these methods with those from the newer Mantel-Haenszel approach, which is also a contingency-table method but one with more statistical power than the three methods studied and easier to compute than the log-linear method. The number and interpretation of items with non-uniform bias would be important in such a comparison.

For the present, this research supported the commonly accepted belief that X2 is an acceptable technique for detecting differential item performance, especially in atypical testing situations. The study also was one of the first to examine the full set of models (main effects and interactions) available with the log-linear method, exploring its ability to look at several variables simultaneously. Such research, it is hoped, will make a contribution to the pursuit of fairness in testing.

APPENDICES

Michigan Educational Assessment Program 1984 - 1985

meap

STATE SUMMARY
Grade 4

OBJ CODE	MATHEMATICS SKILL AREAS AND OBJECTIVES	% Pupils Attained	NUMBER OF PUPILS
	NUMERATION	87	
10-5	ORDER SETS: FEWER	92	104791
10-7	ORDER SETS: FEWEST	92	104686
16-2	PLACE VALUE: HUNDRED CHART	79	104702
16-4	EXPAND 2-DIGIT NUMERAL	83	104748
16-7	EXPAND 3-DIGIT NUMERAL W/WORDS	86	104822
16-8	EXPAND 3-DIGIT NUMERAL	82	104687
16-9	ABC > CBA OR ABC < CBA	93	104855
16-10	ORDER SET OF NUMERALS	88	104822
17-1	NEXT NUMBER IN SEQUENCE	89	104804
	WHOLE NUMBERS	85	
23-1	AB + C, NO REGROUPING	88	104737
23-3	AB + CD, NO REGROUPING	95	104847
24-1	AB + C, WITH REGROUPING	83	104835
24-2	AB + CD, WITH REGROUPING	77	104698
29-2	SUBTRACTION: NUMBER SENTENCE	86	104875
30-1	AB - C, NO REGROUPING	90	104837
30-2	AB - CD, NO REGROUPING	91	104757
31-1	AB - C, WITH REGROUPING	71	104747
35-3	A + A + A... = A x B	81	104788
35-6	A x B = A + A + A...	73	104726
36-1	A x 1 = ?	91	104779
36-3	A x B; A,B < 6	90	104777
	FRACTIONS	87	
79-4/6	IDENTIFY CONGRUENT PARTS	96	104813
79-13	SHADED REGIONS: 1/2, 1/3, 1/4	78	104849
	METRIC MEASUREMENT	85	
107-8	LENGTH: NEAREST CM	79	104876
143-2	TEMPERATURE	92	104861
	NON-METRIC MEASUREMENT	99	
147-6	TIME: NEAREST HOUR	99	104880
	GEOMETRY	95	
156-1	SHAPES	94	104794
163-1	PROPERTIES OF FIGURES	96	104896
	CORRELATED OBJECTIVES	76	
17-3	NUMERATION: ODD OR EVEN	79	104848
28-3	WHOLE NUMBERS: SUBTRACTION	74	104568
29-4	WHOLE NUMBERS: A - B; A,B < 19	79	104573
31-2	WHOLE NUMBERS: AB - CD	58	104626
35-4	WHOLE NUMBERS: 2 x A = ?	72	104538
36-2	WHOLE NUMBERS: A x O = ?	64	104537
48-5)	WHOLE NUMBERS: WORD PROBLEMS	88	104475
156-3	GEOMETRY: SHAPES	94	104578

OBJ CODE	READING SKILL AREAS AND OBJECTIVES	% Pupils Attained	NUMBER OF PUPILS
	VOCABULARY MEANING	88	
IA	PREFIXES	91	104866
IB	SUFFIXES	90	104889
IC	MULTIPLE MEANINGS	91	104841
ID	SYNONYMS	92	104893
IE	ANTONYMS	75	104860
IF	CONTEXT	86	104747
	LITERAL COMPREHENSION	88	
IIB	MAIN IDEA	87	104892
IIC	MAIN IDEA DETAILS	88	104860
IID	SEQUENCE	84	104885
IIF	CAUSE/EFFECT	94	104883
IIG	LIKENESS/DIFFERENCE	87	104883
	INFERENTIAL COMPREHENSION	81	
IIIA	MAIN IDEA	82	104868
IIIB	CAUSE/EFFECT	87	104885
IIIC	PROBABLE OUTCOME	84	104725
IIID	MAIN IDEA DETAILS	72	104887
IIIE	SEQUENCE	80	104879
IIIF	LIKENESS/DIFFERENCE	82	104855
IIIG	CONCLUSIONS	84	104877
IIIH	ANALOGIES	76	104725
IIII	CHARACTERS	86	104859
	CRITICAL READING SKILLS	76	
IVA	AUTHOR'S PURPOSE	76	104738
	RELATED STUDY SKILLS	78	
VA	REFERENCES, AWARENESS	80	104862
VB	REFERENCES, USE	93	104843
VD	SUMMARIZING	76	104879
VF	ALPHABETIZING	65	104805
	POSITIVE RESPONSE/READING	43	
VIA	READ IN FREE TIME	72	104599
VIB	VISIT READING PLACES	26	104554
VIC	REQUEST EXTRA READING	35	104517
VID	TALK ABOUT READING	39	104503
RELATED READING ACTIVITIES -- see Test Item Analysis			

PROPORTIONS REPORT

MATHEMATICS

CATEGORY	1984	1983	1982	1980
4	82.6	80.5	77.8	72.3
3	13.6	14.6	15.9	16.6
2	3.2	3.9	4.7	6.8
1	0.7	1.0	1.5	2.2
Number Of Pupils	104902	107418	112729	134555

STATUS/CHANGE CATEGORY

Status:
Change:

READING

CATEGORY	1984	1983	1982	1980
4	77.7	75.9	75.6	70.7
3	11.9	12.3	12.3	13.6
2	7.1	7.8	7.8	9.6
1	3.3	3.9	4.2	6.1
Number Of Pupils	104914	107471	112779	134555

STATUS/CHANGE CATEGORY

Status:
Change:

TOTAL NUMBER OF STUDENTS INCLUDED IN THIS SUMMARY

104984

APPENDIX B

ITEM BIAS VALUES

Reading

Item No.	C2 (E)	X2 (E)	C2 (S)	X2 (S)	G2	p	Obj No.	
01	0.49	4.38	0.98	6.53	26.41	89	I	A
02	0.74	3.03	0.15	0.83	19.32	88	I	A
03	0.61	3.35	0.36	2.60	18.42	87	I	A
20	8.90	24.22**	2.20	12.71*	45.85**	81	I	B
21	9.64	73.36**	1.05	6.32	91.92**	83	I	B
22	24.71**	98.65**	9.51*	35.36**	140.28**	78	I	B
52	1.54	11.78	0.49	5.28	25.94	88	I	C
53	1.18	14.38	1.05	2.75	24.06	84	I	C
54	0.53	2.42	0.81	6.64	14.79	78	I	C
04	14.80	64.22**	4.45	16.85**	88.55**	84	I	D
05	1.87	11.46	1.41	17.58	36.09**	90	I	D
06	13.48	68.63**	8.96	36.65**	112.77*	75	I	D
37	37.14**	91.92**	13.80**	26.50**	122.29**	60	I	E
38	16.37*	56.39**	3.35	10.85	73.21**	63	I	E
39	9.04	50.48**	3.34	9.85	65.15**	73	I	E
66	1.17	8.27	0.58	2.66	20.78	80	I	F
67	2.54	11.44	0.56	3.63	18.62	79	I	F
68	0.95	14.00	0.54	2.65	17.91	80	I	F
07	3.07	9.46	2.77	7.15	21.67	64	II	B
26	1.23	15.05	0.66	11.35*	34.86	88	II	B
48	1.06	4.63	0.40	2.38	10.34	77	II	B
14	1.40	13.09	0.45	4.65	28.15	91	II	C
42	2.01	8.39	3.42	12.53*	35.91	75	II	C
63	5.75	14.84	0.52	4.51	35.10	78	II	C
18	2.28	9.76	0.97	5.91	22.93	82	II	E
30	3.28	11.62	0.83	4.00	23.21	69	II	E
46	2.43	8.54	8.85	25.16**	36.60	74	II	E
10	0.73	8.95	0.89	5.82	33.80	91	II	F
32	4.20	17.38	1.34	4.06	23.21		II	F
58	1.48	13.53	2.57	8.77	27.47	89	II	F
09	6.25	21.74*	2.53	6.91	48.09*	75	II	G
28	3.67	10.19	0.20	1.67	20.61	80	II	G
50	3.31	20.85*	0.89	5.53	39.39*	79	II	G

* = $p < .05$ ** = $p < .01$

C2 = Scheuneman C2 E = Ethnic
X2 = full X2 S = Sex
G2 = log-linear

Note: Not all objectives are measured on every test: e.g., II A was not used on this test.

Reading (continued)

Item No.	C2 (E)	X2 (E)	C2 (S)	X2 (S)	G2	p	Obj No.
13	1.81	8.09	0.76	2.69	17.25	79	III A
41	3.81	20.09*	2.53	6.27	29.65	72	III A
62	5.18	14.72	1.62	3.52	31.31	63	III A
17	4.39	17.28	9.74*	60.73**	88.79*	84	III B
29	11.47	28.07**	0.87	1.99	44.34*	67	III B
45	0.70	7.29	13.73**	51.62**	64.51*	80	III B
40	0.74	7.56	1.63	9.66	28.71	82	III C
73	3.53	18.84*	1.47	10.13	45.93*	72	III C
75	2.87	6.57	5.59	12.56*	34.62	74	III C
08	12.52	25.32**	13.69**	22.19**	56.12**	45	III D
27	7.85	26.64**	1.06	5.56	39.68*	67	III D
49	2.73	9.55	0.98	3.97	24.07	67	III D
11	5.44	24.09**	1.42	4.08	37.90*	75	III E
33	2.15	9.14	1.97	7.16	24.20	70	III E
59	1.02	2.21	4.45	7.32	15.60	66	III E
15	20.34**	45.62**	13.07**	30.42**	101.23**	43	III F
43	2.69	14.17	0.19	2.63	23.36	80	III F
64	1.09	7.58	0.54	3.16	14.56	77	III F
12	4.43	15.19	0.59	1.59	25.19	70	III G
34	3.78	18.50*	1.60	7.12	37.39	76	III G
60	1.87	10.42	0.10	1.04	22.12	76	III G
69	3.31	7.81	1.47	3.41	22.72	57	III H
70	9.88	45.34**	3.74	13.91*	68.56*	70	III H
71	1.63	6.29	1.35	6.51	17.45	76	III H
16	4.43	16.05	2.54	9.45	42.55*	81	III I
44	1.78	8.11	1.97	4.77	29.48*	70	III I
65	0.22	7.48	0.52	4.51	24.25	80	III I
61	2.95	7.39	2.92	8.43	33.02*	60	IV A
72	4.13	11.24	1.91	9.76	54.44**	66	IV A
74	6.84	26.76**	1.06	3.29	46.98**	71	IV A
23	13.89	30.83**	0.59	1.99	48.32**	66	V A
24	2.61	10.73	1.92	5.63	21.43	63	V A
25	1.28	9.83	0.24	1.53	26.19	83	V A
35	1.06	12.26	0.89	5.79	34.98	91	V B
36	1.08	13.42	0.46	7.07	30.31	93	V B
51	4.90	10.92	2.53	14.62*	30.11	69	V B
19	7.31	15.13	6.99	12.35*	35.45	49	V D
31	1.97	6.60	2.51	6.64	19.04	68	V D
47	2.80	19.05*	4.88	16.49**	54.15*	74	V D
55	5.54	18.21	4.35	10.39	39.29*	58	V F
56	5.25	13.05	1.60	6.88	29.56	60	V F
57	4.64	11.72	8.91	14.50*	34.42	58	V F

Mathematics

Item No.	C2 (E)	X2 (E)	C2 (S)	X2 (S)	G2	p	Obj No.
133	2.79	21.64*	0.67	8.95	41.01*	90	10-5
134	1.94	11.73	0.86	7.56	31.95	90	10-5
135	1.19	11.06	0.95	8.55	29.26	90	10-5
169	3.23	26.87**	0.16	1.21	41.12*	88	10-7
170	1.90	17.74	0.44	6.39	34.54	89	10-7
171	3.31	22.80*	0.19	2.49	42.83*	88	10-7
97	3.53	20.28*	1.91	10.93	39.64*	81	16-2
98	2.61	18.18	0.85	5.40	36.28	83	16-2
99	2.55	21.18*	0.23	2.58	32.34	84	16-2
151	11.30	31.47*	4.37	22.78*	58.27*	75	16-4
152	4.42	13.89	4.18	27.48*	45.27*	79	16-4
153	2.79	9.63	5.07	15.38*	32.20	73	16-4
121	11.74	27.74**	1.44	8.36	49.31**	74	16-7
122	2.13	11.19	1.19	10.06	44.01*	83	16-7
123	1.23	6.95	1.81	6.11	35.36*	88	16-7
163	5.12	12.25	0.65	5.08	37.99*	79	16-8
164	2.52	7.66	2.37	14.12*	46.11*	81	16-8
165	10.45	25.73**	0.60	3.12	52.44**	73	16-8
112	0.92	9.65	0.95	13.66*	42.73*	93	16-9
113	0.99	11.48	0.43	5.96	34.14	92	16-9
114	0.40	6.54	0.42	4.54	24.56	92	16-9
109	2.42	11.41	0.92	7.46	29.40	84	16-10
110	1.92	19.31	0.51	4.55	29.87	89	16-10
111	3.58	22.67*	0.68	4.07	41.58*	86	16-10
124	1.55	11.91	0.52	6.81	29.28	90	17-1
125	2.58	9.03	0.26	1.38	25.52	81	17-1
126	1.04	9.21	0.74	4.52	23.09	87	17-1
154	1.77	16.82	1.67	7.99	36.79	86	23-1
155	4.03	21.54*	3.73	18.53**	48.99*	87	23-1
156	1.44	11.20	1.13	10.40	36.56	90	23-1
115	0.93	15.05	0.34	5.97	32.66	94	23-3
116	0.89	10.90	0.79	4.80	30.05	91	23-3
117	1.04	13.19	0.12	1.36	32.63	93	23-3
118	4.65	16.75	1.57	6.93	30.75	77	24-1
119	5.09	16.57	1.44	13.13*	41.83*	78	24-1
120	2.16	15.24	0.19	2.12	34.30*	83	24-1
160	6.40	18.35*	9.55*	26.99**	53.72**	70	24-2
161	3.30	12.83	14.71**	45.90**	65.54**	71	24-2
162	5.35	14.70	9.35	27.39**	58.45*	70	24-2
100	1.30	7.87	1.96	14.16*	30.60	85	29-2
101	1.93	13.14	1.14	7.25	29.47	87	29-2
102	1.98	14.10	0.95	7.24	30.60	87	29-2
166	0.49	6.09	2.50	17.09**	33.14	87	30-1
167	2.39	17.82	2.21	13.39*	45.95**	87	30-1
168	3.39	17.07	2.72	17.38**	47.72**	88	30-1

Mathematics (continued)

Item No.	C2 (E)	X2 (E)	C2 (S)	X2 (S)	G2	p	Obj. No.
148	4.05	15.57	3.98	20.32**	55.14*	86	30-2
149	2.78	11.48	2.77	18.01**	45.77**	87	30-2
150	0.92	8.04	2.65	18.79**	38.09*	91	30-2
145	1.49	5.14	3.68	7.20	24.89	66	31-1
146	2.25	15.47	1.50	5.68	26.52	68	31-1
147	3.74	10.15	1.29	5.11	23.43	68	31-1
130	0.85	11.74	0.26	1.68	31.64*	88	35-3
131	4.76	24.18**	2.34	11.70*	50.01**	76	35-3
132	1.24	9.08	0.15	0.83	19.23	77	35-3
157	6.88	17.21	5.81	14.60*	46.79**	67	35-6
158	8.28	24.96**	3.43	8.66	44.25*	66	35-6
159	6.94	27.59**	7.76	24.95**	65.52**	75	35-6
139	2.81	25.35	0.78	8.69	43.19*	92	36-1
140	1.50	10.88	0.15	2.36	31.76	89	36-1
141	4.01	18.03	0.42	3.75	36.03	89	36-1
142	3.54	13.63	0.96	9.59	33.77	86	36-3
143	3.06	13.12	0.29	4.75	34.45	88	36-3
144	3.26	15.84	0.06	0.46	22.20	87	36-3
136	0.89	10.91	0.94	13.62*	36.21	92	79-4/6
137	1.91	22.72*	2.00	14.32*	41.65*	91	79-4/6
138	2.56	25.80**	0.77	8.09	50.48**	95	79-4/6
106	4.69	17.99	0.53	3.46	27.74	78	79-13
107	4.98	34.70**	0.66	2.97	52.80*	83	79-13
108	4.11	17.03	3.62	10.39	41.94*	79	79-13
91	4.98	16.27	2.68	14.00*	38.34*	77	107-8
92	1.42	4.77	2.73	11.07*	18.52	75	107-8
93	2.32	7.11	5.08	20.61**	29.80	74	107-8
103	3.36	20.43*	7.30	27.95**	59.24*	84	143-2
104	2.43	15.36	1.54	6.24	42.32*	87	143-2
105	0.65	14.52	0.45	9.15	32.79	96	143-2
94	0.22	11.74	0.08	9.09	31.57	98	147-6
95	0.13	11.78	0.04	6.45	26.61	99	147-6
96	0.23	20.42	0.02	2.41	28.16	99	147-6
127	1.18	16.77	0.47	4.02	35.40	94	156-1
128	1.75	15.02	0.19	2.11	33.64	88	156-1
129	3.15	21.76*	0.89	11.36*	42.81*	90	156-1
172	1.83	18.87	0.75	10.90	44.46*	95	163-1
173	0.46	12.21	0.26	10.56	35.13*	95	163-1
174	1.75	19.89*	0.46	5.21	36.24	94	163-1

LIST OF REFERENCES

LIST OF REFERENCES

Alderman, D. L., & Holland, P. W. (1980). Item performance across native language groups on the Test of English as a Foreign Language (Research Report 81-16). Princeton, NJ: Educational Testing Service.

Andrews, F. M., Klein, L., Davidson, T. N., O'Malley, P. M., & Rodgers, W. L. (1981). A guide for selecting statistical techniques for analyzing social science data. Ann Arbor, MI: Institute for Social Research.

Angoff, W. H. (1972). A technique for the investigation of cultural differences. Paper presented at the annual meeting of the American Psychological Association, Honolulu.

Angoff, W. H. (1982). Use of difficulty and discrimination indices for detecting item bias. In R. A. Berk (Ed.), Handbook of methods for detecting test bias (pp. 96-116). Baltimore: Johns Hopkins University.

Angoff, W. H., & Ford, S. F. (1973). Item-race interaction on a test of scholastic aptitude. Journal of Educational Measurement, 10, 95-105.

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

Baker, F. B. (1981b). Log-linear, logit-linear models: A didactic. Journal of Educational Statistics, 6, 75-102.

Beck, M. D., & Sklar, J. (1978). An assessment of item "bias" in the 1978 Metropolitan Achievement Tests. Paper presented at the annual meeting of the Eastern Educational Research Association, Williamsburg, VA.

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

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). Discrete multivariate analysis: Theory and practice. Cambridge, MA: MIT Press.

Bock, R. D. (1975). Multivariate statistical methods in behavioral research. New York: McGraw-Hill.

Bock, R. D., & Yates, G. (1973). MULTIQUAL: Log-linear analysis of nominal or ordinal qualitative data by the method of maximum likelihood. Chicago: National Educational Resources, Inc.

Burrill, L. E. (1982). Comparative studies of item bias methods. In R. A. Berk (Ed.), Handbook of methods for detecting test bias. Baltimore, MD: Johns Hopkins University.

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

Census Bureau. (1983). 1980 Census of Population. Vol. I: Characteristics of the population. Chpater C: General social and economic characteristics. Part 24: Michigan. (PC80-1-C24). Washington, DC: U.S. Government Printing Office.

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

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

Cohen, J. (1969). Statistical power analysis for the behavioral sciences. New York: Academic Press.

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

Cronbach, L. J. (1975). Five decades of public controversy over mental testing. American Psychologist, 30, 1-14.

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

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. Journal of Educational Measurement, 23, 355-368.

Durovic, J. J. (1975). Test bias: An objective definition for test items. Paper presented at the annual meeting of the Northeastern Educational Research Association, Ellenville, NY. (ED 128 381)

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

Educational Testing Service. (1980). An approach for identifying and minimizing bias in standardized tests: A set of guidelines. Princeton, NJ: Office for Minority Education.

Einhorn, H. J., & Bass, A. R. (1971). Methodological considerations relevant to discrimination in employment testing. Psychological Bulletin, 75, 261-269.

Feinberg, S. E. (1977). The analysis of cross-classified categorical data. Cambridge, MA: MIT Press.

Frary, R. B., & Zimmerman, D. W. (1984). Elimination of bias in test scores: Effect on reliability and validity. Educational and Psychological Measurement, 44, 25-31.

Friedman, C. B. (1984). The construct validation of second language proficiency tests with different native language groups. Unpublished doctoral dissertation, Indiana University-Bloomington.

Green, D. R., & Draper, J. F. (1972). Exploratory studies of bias in achievement tests. Paper presented at the annual meeting of the American Psychological Association, Honolulu. (ED 070 794)

Harris, D. J., & Hoover, H. D. (1986). The stability of selected item bias indices. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Holland, P. W., & Thayer, D. T. (1986). Differential item performance and the Mantel-Haenszel procedure. Paper presented at the annual meeting of the American Educational Research Association, San Francisco. (ED 272 577)

Hoover, H. D., & Kolen, M. J. (1984). The reliability of six item bias indices. Applied Psychological Measurement, 8, 173-181.

Hoover, M. E. R. (1984). Teacher competency tests as educational genocide for blacks: The Florida Teacher Certification Examination. Negro Educational Review, 35, 70-77.

Hunter, J. E. (1975). 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 Conference on Test Bias, Annapolis, MD.

Hunter, R. V., & Slaughter, C. D. (1980). ETS test sensitivity review process. Princeton, NJ: Educational Testing Service.

Ironson, G. H. (1982a). Past accomplishments, future needs. Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Ironson, G. H. (1982b). Use of chi-square and latent trait approaches for detecting item bias. In R. A. Berk (Ed.), Handbook of methods for detecting test bias (pp. 117-160). Baltimore: Johns Hopkins University.

Ironson, G. H., & Craig, R. (1982). Item bias when amount of bias is varied and score differences between groups are present (Final Report NIE-G-81-0045). Tampa: University of South Florida. (ED 227 146)

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

Ironson, G. H., Homan, S., Willis, R., & Signer, B. (1984). The validity of item bias techniques with math word problems. Applied Psychological Measurement, 8, 391-396.

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

Kennedy, J. J. (1983). Analyzing qualitative data: Introductory log-linear analysis for behavioral research. New York: Praeger.

Kok, F. G., Mellenbergh, G. J., & van der Flier, H. (1985). Detecting experimentally induced item bias using the iterative item bias logit method. Journal of Educational Measurement, 22, 295-303.

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

Lippman, W. (1986). The great confusion. Educational Forum, 50, 371-374. (Reprinted from the New Republic, 1923.)

Loyd, B. H. (1984). Evaluation of log linear models for detection of item bias: A comparison across samples. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

Loyd, B. H. (1985). Detection of item bias: A comparison across methods. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Loyd, B. H. (1986). Differential item performance: The interaction of method and content. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Macmillan Publishing Company. (1975). Guidelines for creating positive sexual and racial images in educational materials. New York: Author.

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

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. Journal of Educational Statistics, 7, 105-118.

Merz, W. R., & Grossen, N. (1979). An empirical investigation of six methods for examining item bias. Paper presented at the annual meeting of the National Council on Measurement in Education. (ED 178 566)

Office of Management and Budget. (1979). Race and ethnic standards for federal statistics and administrative reporting (OMB Directive No. 15). Washington, DC: U.S. Government Printing Office.

Pennock-Roman, M. (1983). An application of the Del statistic to detect differential item performance on the Test of English as a Foreign Language. Paper presented at the annual meeting of the American Educational Research Association, Montreal.

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

Phelps, J. L., Donovan, D. L., Roeber, E. D., Carr, R. A., & Caswell, M. S. (1980). Technical report. Vol. 1. Michigan Educational Assessment Program. Lansing, MI: Michigan State Board of Education.

Phelps, J. L., Donovan, D. L., Roeber, E. D., Carr, R. A., & Caswell, M. S. (1981). Technical report. Vol. 2. Michigan Educational Assessment Program. Lansing, MI: Michigan State Board of Education.

Raju, N. S., & Normand, J. (1985). The regression bias method: A unified approach for detecting item bias and selection bias. Educational and Psychological Measurement, 45, 37-54.

Reynolds, C. R., & Brown, R. T. (Eds.) (1984). Perspectives on bias in mental testing. New York: Plenum Press.

Roeber, E. D. (1984). MEAP Hispanic coding study. Lansing, MI: Michigan State Department of Education.

Rudner, L. M., & Convey, J. J. (1978). An evaluation of select approaches for biased item identification. Paper presented at the annual meeting of the American Educational Research Association, Toronto. (ED 157 942)

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

Scheuneman, J. D. (1975). A new method of assessing bias in test items. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC. (ED 106 359)

Scheuneman, J. D. (1976). Validating a procedure for assessing bias in test items in the absence of an outside criterion. Paper presented at the annual meeting of the American Educational Research Association, San Francisco. (ED 129 853)

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

Scheuneman, J. D. (1980). Latent-trait theory and item bias. In L. J. Th. van der Kamp (Ed.), Psychometrics for educational debates (pp. 139-151). New York: John Wiley & Sons.

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

Scheuneman, J. D. (1982a). A posteriori analyses of biased items. In R. A. Berk, (Ed.), Handbook of methods for detecting test bias (pp. 180-198). Baltimore: Johns Hopkins University.

Scheuneman, J. D. (1982b). Item bias and test scores. Paper presented at the annual meeting of the National Council for Measurement in Education, New York. (ED 219 450)

Scheuneman, J. D. (1987). An experimental, exploratory study of causes of bias in test items. Journal of Educational Measurement, 24, 97-118.

Schmeiser, C. B. (1982). Use of experimental design in statistical item bias studies. In R. A. Berk (Ed.), Handbook of methods for detecting test bias. Baltimore, MD: Johns Hopkins University.

Schmeiser, C. B. (1985). "Debiasing" standardized tests: A case study. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

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

Shepard, L., Camilli, G., & Williams, D. M. (1984). Accounting for statistical artifacts in item bias research. Journal of Educational Statistics, 9, 93-128.

Shepard, L., Camilli, G., & Williams, D. M. (1985). Validity of approximation techniques for detecting item bias. Journal of Educational Measurement, 22, 77-105.

Stricker, L. (1982). Identifying test items that perform differentially in population subgroups: A partial correlation index. Applied Psychological Measurement, 6, 261-273.

Subkoviak, M. J., Mack, J. S., Ironson, G. H., & Craig, R. D. (1984). Empirical comparison of selected item bias detection procedures with bias manipulation. Journal of Educational Measurement, 21, 49-58.

Terman, L. M. (1986). The great conspiracy. Educational Forum, 50, 363-370. (Reprinted from the New Republic, 1922.)

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

Tittle, C. K. (1982). Use of judgmental methods in item bias studies. In R. A. Berk (Ed.), Handbook of methods for detecting item bias (pp. 31-63). Baltimore: Johns Hopkins University.

van der Flier, H., Mellenbergh, G., Ader, H. J., & Wijn, M. (1984). An iterative item bias detection method. Journal of Educational Measurement, 21, 131-145.

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

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