DOCUMENT RESUME

ED 340 735 TM 017 792

AUTHOR O'Neal, Marcia R.
TITLE A Comparison of Methods for Detecting Item Bias.
PUB DATE Nov 91
NOTE 27p.; Paper presented at the Annual Meeting of the Mid-South Educational Research Association (20th, Lexington, KY, November 12-15, 1991).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Black Students; Comparative Analysis; Correlation; Difficulty Level; Females; Grade 9; High Schools; *High School Students; *Item Bias; *Language Tests; Males; *Mathematics Tests; Minimum Competency Testing; Pilot Projects; *Reading Tests; Test Construction; *Test Items; White Students
IDENTIFIERS *Alabama Basic Competency Tests; Angoff Methods; Delta Plot; Distractor Technique; Mantel Haenszel Procedure; Procedure (Angoff)

ABSTRACT

Six item bias detection techniques (distractor analysis, the Mantel-Haenszel procedure, Angoff's delta plot, Angoff's modified delta plot, Stricker's partial correlation index, and direct comparison of item difficulties) were compared for their ability to identify biased items at three stages of test development. The techniques were applied to items administered during the spring 1989 item pilot (n=about 1,000 for each of 36 forms), the fall 1989 form pilot (n=about 460 for each of three tests), and the spring 1990 final test version (n=about 46,000 each for the reading, mathematics, and language tests) of the Grade 9 Alabama Basic Competency Tests. Four subgroup pairs were analyzed for the reading, mathematics, and language tests; each administration; and each technique. The subgroup pairs included black versus white females, black versus white males, black females versus black males, and white females versus white males. The method with the greatest stability across administrations was the direct difficulty difference, followed closely by distractor analysis. The Mantel-Haenszel procedure and the partial correlation index both showed weak to moderate stability. The poorest results concerning stability were found for both Angoff techniques. The modified Angoff version is not a desirable technique, since it can identify inappropriate items. Results suggest the possible use of more than one index to examine a set of items. A 60-item list of references is included. (SLD)

A COMPARISON OF METHODS FOR DETECTING ITEM BIAS

Marcia R. O'Neal
Evaluation and Assessment Laboratory
The University of Alabama

Paper presented at the annual meeting of the Mid-South Educational Research Association, Lexington, Kentucky, November 13-15, 1991.

A COMPARISON OF METHODS FOR DETECTING ITEM BIAS

Introduction

Tests and testing play a major role in today's society. It is vital, therefore, that test developers and users strive to ensure the validity of their instruments for the populations and purposes for which they are intended. Toward this end, many writers and investigators have addressed the issue of test bias. Earliest discussions of test bias date back to the first decade of the twentieth century (Eells, Davis, Havighurst, Herrick, & Tyler, 1951; Jensen, 1980; Osterlind, 1983). As Scheuneman (1981) indicated, however, increased interest in test bias was stimulated by the civil rights movement in the late 1960s, and the subject has been prevalent in the literature on measurement since that time.

Test bias clearly will remain an issue of considerable importance as well as one requiring the attention of measurement specialists and others, not only because of recognized professional responsibilities, but also because of criticism and litigation directed at tests and testing. Court cases repeatedly have focused attention on the issue of bias in testing in a variety of fields. For example, the case of Deborah P. vs. Turlington challenged Florida's competency testing program in the schools partly on the basis of racial discrimination (Overcast & Sales, 1982). In the case of the Golden Rule Life Insurance Company vs. Mathias, the charge involved racial bias in examinations for licensing Illinois insurance agents (Faggen, 1987; Rooney, 1987). Both the longstanding Larry P. vs. Riles case in California involving allegations of bias in IQ tests and the more recent Allen vs. the Alabama Board of Education litigation directed against the teacher certification test have been cited by the Fairtest Examiner, a publication of the National Center for Fair and Open Testing ("I.Q. Tests Banned," 1987; "Teacher Testing," 1987).

A major component of the test bias issue is the question of whether the items that comprise the test are biased. As with the larger issue, item bias was first addressed in the early 1900s by individuals such as Alfred Binet and William Stern (Eells et al., 1951; Jensen, 1980). Jensen (1980) and Osterlind (1983) both recognized the 1951 study reported by Eells et al. (1951) as the first important advance following those early pioneers in item bias research, declaring the Eells study a classic in the literature.

The procedures in use today for examining item bias were developed more recently, making item bias research a very young field (Osterlind, 1983).

Definitions of Item Bias

Many definitions of item bias have been offered. Rudner (1978b) suggested a concise definition, stating that a biased item is one that "behaves differently for members of two different culture groups" (p. 33). Shepard, Camilli, and Averill (1981) stated that biased items are items that are found to be "deviant or anomalous . . . in the context of other items. . . . [They] may be measuring different things for different groups" (p. 317). Jensen (1980) viewed item bias detection within the framework of internal detection of test bias. He offered the following comments:

> Most tests are composed of a number of items that singly and in relationship to one another have a variety of statistical properties that can be compared across different populations. If certain of these statistical properties of the test differ significantly in any two populations, it is prima facie evidence that the test internally behaves differently in the two populations and one may suspect that the test is biased with respect to these particular populations. (p. 429)

A content validity definition was provided by Reynolds (1982b).

> An item or subscale of a test is considered to be biased in content when it is demonstrated to be relatively more difficult for members of one group than another when the general ability level of the groups being compared is held constant and no reasonable theoretical rationale exists to explain group differences on the item (or subscale) in question. (p. 188)

The AERA, APA, NCME joint committee developed the following definition of item bias which was published in the 1985 Standards for Educational and Psychological Testing:

> An item is considered positively or negatively biased for a group within a population if (a) the average expected item score for that group is substantially higher or lower than that for the overall population and (b) if this disparity stems from factors that the item is not intended to measure rather than from factors it is intended to measure. (p. 92)

Implied in several of these definitions, notably those of Shepard et al. (1981) and Jensen (1980), is the notion that empirical indices of bias do not necessarily indicate bias in the sense that it means unfairness. As Angoff (1982) indicated, "there should be an educational and psychological rationale for deciding that a statistically biased item is indeed biased" (p. 114). Further, Angoff (1988) indicated that

bias indices "only tell us that the item may be more difficult for one of the groups being studied," and that a finding of bias is "a matter for the investigator to decide" (p. 215). Likewise, Osterlind (1983) stated:

> The term [bias] is conceptually distinct and operationally different from the concepts of fairness, equality, prejudice, or preference or any of the other connotations sometimes associated with its use in popular speech. Bias, then, is a technical term and denotes nothing more or less than the consistent distortion of a statistic. (pp. 10-11)

Typically definitions of item bias are tied to specific techniques or classes of techniques. Scheuneman (1981) suggested that most definitions can be found to fall into one of two general categories: those concerned with item-by-group interaction and those contingent on ability levels. Mellenbergh (1981) referred to these definitional categories as unconditional and conditional. Rudner's (1978b) definition would be an example of the first of the two Scheuneman categories, while Reynolds' (1982b) would fall under the second category.

Item Bias Detection Techniques

The numerous empirical techniques available for detecting item bias may be categorized in several ways. Borrowing from Crocker and Algina (1986), Jensen (1980), Osterlind (1983), Peterson (1980), and Shepard (1982), a classification scheme may be used that groups item bias detection techniques into one of six different categories. The groups include analysis of variance approaches, factor analytic techniques, distractor analysis, indices based on item difficulty or item discrimination, item characteristic curve methods, and chi-square techniques.

Using the analysis of variance approach, evidence of bias is investigated through the use of a groups-by-items repeated measures ANOVA design which uses item p values or transformations of them. Of primary interest in this type of study is the groups-by-items interaction effect (Angoff, 1982; Crocker & Algina, 1986; Jensen, 1984; Osterlind, 1983; Peterson, 1980). With an ANOVA approach, an item is considered biased if the difference in group means for the item difficulty level is unusually high or low relative to other items (Crocker & Algina, 1986; Peterson, 1980).

Factor analysis is actually a group of sophisticated statistical techniques that can be employed to examine the construct validity of tests. It provides a method for examining the intercorrelations among variables or elements and then combining related elements in such a way that fewer variables (called

factors) are needed to account for the intercorrelations. Factor analysis is a way of describing behavior in simpler terms by using fewer categories than in the original set of elements (Anastasi, 1982). The use of factor analytic techniques to detect item bias involves comparing factor analysis results for the groups in question to see if consistency exists across groups. The intent in using factor analysis for item bias research is to determine which items, if any, are not constant across groups. These are the items identified as biased (Reynolds, 1982a).

One method reported infrequently in the literature on item bias methods is distractor analysis. In this procedure, incorrect response alternatives, called distractors or question foils, are examined to determine if different groups of examinees exhibit different patterns of responses. Unbiased items are those whose distractors have the same relative attractiveness for each group (Jensen, 1980; Osterlind, 1983; Peterson, 1980). The rationale for the procedure was stated by Veale and Foreman (1983), who originally proposed the strategy in 1975.

> The approach . . . is based on the notion that examinees' responses to the incorrect options of a multiple-choice item, called distractors or foils, provide more and better information concerning cultural bias than their responses to the correct option. When one group is attracted to a particular foil and the other groups are drawn to other foils, cultural bias may be present in the item. This response configuration will likely occur when (1) a foil is actually the correct response for a particular cultural group or (2) a foil, although clearly incorrect for all groups, contains culture-specific stimuli which attract (or repel) members of some group. The proposed method also indicates the likely source of the bias so that the item may be revised to eliminate the bias rather than discarding the item. (p. 249)

Several methods for detecting item bias have been designed to focus on examination of differences between groups with respect to item difficulty or item discrimination values. In general, item difficulty is a percentage value representing the proportion of examinees who answer an item correctly. Thus, an item difficulty or p value of .60 indicates that 60 percent of those taking the item gave a correct response. Higher p values indicate easier items, and lower p values indicate more difficult items. One of the most widely known techniques falling into the category of item difficulty is the delta-plot method described by Angoff (1972; Angoff & Ford, 1973; Angoff & Sharon, 1974). The procedure, sometimes referred to as transformed item difficulty, or TID (Osterlind, 1983), is straightforward. Item difficulties for the two groups are converted to delta values (normal deviates), and these values are correlated and plotted on a

bivariate graph (Angoff, 1982; Crocker & Algina, 1986; Jensen, 1980). A straight line can be fitted to the scatterplot, and the perpendicular distances of the points from the line can be computed. Large distances of the plotted points from the fitted line are indicators of bias.

Item discrimination is a measure of the extent to which an item distinguishes between those examinees possessing "more" of the trait being measured and those who possess "less" of that trait. According to Jensen (1980), item discrimination is "the correlation of each item with the total score on the test, which indicates to what extent a particular item measures whatever is measured by the test as a whole" (p. 137). A number of indices have been developed to deal with differing specific situations in which a measure of item discrimination is needed. Berk (1984) identified four preferred and thirteen less practical indices. Anastasi (1982) indicated that, despite the variety of indices, the results have been found to be quite similar in terms of which items are identified as biased.

Methods using the item characteristic curve (ICC) as a basis for determining item bias are based on item response theory (IRT) (Osterlind, 1983). The basic tenet of this theory is that an underlying trait possessed by individuals accounts for the way they respond to an item (Crocker & Algina, 1986; Osterlind, 1983; Peterson, 1980). The mathematical model for IRT describes the response patterns for individuals at differing ability levels. What makes IRT so useful in item bias research is that groups of differing ability levels can be compared (Crocker & Algina, 1986). The ICC "plots the probability of responding correctly to an item as a function of the latent trait (denoted by θ) underlying performance on the items on the test" (Crocker & Algina, 1986, p. 340). The general procedure for using ICCs for item bias research is to produce an ICC for each item for each subgroup. The ICCs of each subgroup are then compared for a given item. An unbiased item is one that results in identical ICCs for all groups being considered (Ironson, 1982). The definition of bias using this approach is conditional on ability.

Chi-square methods, like the IRT approaches, are based on conditional definitions of bias. Chi-square methods involve the examination of differences between observed and expected responses using a goodness-of-fit test after individuals in the two differing culture groups have been divided into categories based on their ability levels as measured by their total test scores (Osterlind, 1983). An item is considered

unbiased if the proportion of correct responses is the same for the two groups (Rudner et al., 1980a). Chi-square procedures may be considered an approximation of IRT approaches (Crocker & Algina, 1986) in that they are actually comparing two item response curves (Peterson, 1980). Crocker and Algina (1986) noted that chi-square methods are much simpler to apply than IRT procedures.

Findings from Previous Comparison Studies

A number of investigators have sought to compare various sets of item bias methods (Burrill, 1981; Devine & Raju, 1982; Hambleton & Rogers, 1989; Hoover & Kolen, 1984; Ironasuwan, 1979; Ironson & Craig, 1982; Ironson, Homan, Willis, & Signer, 1984; Ironson & Subkoviak, 1979; Laksana, 1979; Merz & Grossen, 1978; Nungester, 1977; Perlman, Bezruczko, Junker, Reynolds, Rice, & Schulz, 1988; Phillips & Mehrens, 1988; Raju & Normand, 1985; Rudner, 1977/1978a; Rudner, Getson, & Knight, 1980b; Scong & Subkoviak, 1987; Shepard, Camilli, & Averill, 1981; Shepard, Camilli, & Williams, 1985; Stricker, 1982; Stricker, 1984; Subkoviak, Mack, Ironson, & Craig, 1984; Welch, Ackerman, Doolittle, & Hurley, 1987; Wise, 1987). Some have used real data while others have used simulated data for the purpose of comparing techniques.

The results of these comparison studies, taken as a group, are mixed. While some investigators have attested to the comparability of various techniques, others have found a lack of agreement among the bias indices. The use of simulated data and the control of variables in ways indicated by some investigators have helped to identify in a few instances the conditions under which some of the methods may be preferred. For example, the insensitivity of transformed item difficulties to variations in discrimination values suggests a modification of the method or the selection of another method when it is known or suspected that such values vary widely. Small sample sizes make IRT techniques less appropriate. In other instances, however, results among the studies are inconclusive or contradictory.

What has emerged from an evaluation of these studies is the fact that several promising or appealing techniques have not been studied adequately and others have not been compared at all. For example, none of the comparison studies examined Veale and Foreman's (1983) distractor analysis or Osterlind's (1983) suggested modification, and only two studies (both by Stricker) used Stricker's (1981,

1982) partial correlation index (an item difficulty approach designed to control for ability), despite the fact that both Angoff (1988) and Osterlind (1983) have indicated the potential for these techniques if evaluations of them prove fruitful. Noticeably absent from the list of techniques compared is a controversial item difficulty approach known as the Golden Rule procedure. It has been summarily rejected in the professional literature (Denton, 1988; Rebell, 1988), but writers have indicated increasing legislative efforts to make use of the method. It would therefore seem valuable to establish its performance relative to other more accepted methods. A method that has received inadequate study to date is a modified delta plot procedure recommended by Angoff (1982) as a solution to the problem of insensitivity of the delta plot method to item discrimination differences. This modification has received attention in only one comparison study. Likewise, the Mantel-Haenszel technique (Mantel & Haenszel, 1959), a member of the chi-square family, while receiving some recent attention, needs further study.

Another gap revealed by an examination of these comparison studies is the inadequate representation of criterion-referenced tests, and more particularly, minimum competency tests among the comparisons. Most evaluations involved norm-referenced instruments, Monte Carlo studies, or contrived data sets. Of the 24 studies examined, 12 involved norm-referenced instruments and 7 involved either simulated or contrived data sets. Two studies included both norm-referenced tests and simulated data. Only three comparison studies applied item bias methods to criterion-referenced tests.

Finally, only one study was found that examined items at the item pilot or tryout phase of test development, and few studies were found that measured the stability of item bias indices. Those that have done so typically have taken samples from the same pool of examinees so that the techniques are being compared for items administered in the same context for each comparison.

Purpose of the Study

The purpose of this study was to investigate and compare six item bias techniques based on their ability to identify items at three stages of test development. The six techniques examined included distractor analysis, one chi-square technique (the Mantel-Haenszel procedure), and four item difficulty approaches (Angoff's delta plot, Angoff's modified delta plot, Stricker's partial correlation index, and

direct comparison of item difficulties). These indices were selected based on the potential for their use as suggested by the literature, their limited representation in comparison studies, and/or their acceptability for use with small samples. These techniques were applied to items administered during the item pilot, form pilot, and final test version of the Grade 9 Alabama Basic Competency Tests (BCT). Four subgroup pairs were analyzed for each test (reading, mathematics, and language), each administration, and each technique. The subgroup pairs included black versus white females, black versus white males, black females versus black males, and white females versus white males.

Research Questions

1. What items are identified as biased by each of the six methods when they are applied to the item pilot, form pilot, and final form of the BCT?

2. How and to what extent do the six methods being compared agree in their identification of biased items when they are applied to the item pilot, form pilot, and final form of the BCT?

3. To what extent does each method predict item bias on the form pilot and final form based on the method's identification of items on the item pilot, and on the final form based on the method's identification of items on the form pilot?

Method

Techniques Compared

Distractor Analysis: Using Osterlind's (1983) suggestion and description of the procedure, all responses, both incorrect and correct, are included in the analysis. The first step is to obtain each item's response choice distribution for each of the subgroups. These data are placed into a set of 2-by-K contingency tables where K represents the number of subgroups being compared, and the number of tables per item is equal to the number of answer choices available. Next a chi-square test of independence is computed for each response alternative for the item. A response alternative is unbiased if the chi-square value obtained does not exceed the critical chi-square value. For the present analysis, the highest chi-square value among the four values (representing four response choices) computed for a given item was selected for inclusion in the comparisons. When the analyses involve 2-by-2 tables, as was true in the

present analysis, a simplified formula can be used (Ferguson, 1981). The formula used for this study included Yates' correction for continuity.

Angoff's Delta Plot: This method requires that item difficulties (p values) first be computed for each subgroup. These p values are then converted to z values. Next the z values are converted to deltas by the formula $\Delta = 4z + 13$. At this point, pairs of deltas (each representing one of two subgroups) for each item can be plotted on a bivariate graph. The formulas used in the present study for obtaining the slope, intercept, and perpendicular distance of each point from the major axis are those provided by Angoff and Ford (1973). Greater perpendicular distances represent greater degrees of bias.

Angoff's Modified Delta Plot: The modified delta plot suggested by Angoff (1982) follows the same procedures as the delta plot described above, except that the z values are divided by the correlation of the item with the total test. The resulting value is denoted by z'. Deltas are then obtained by the formula $\Delta' = 4z' + 13$, and the slope, intercept, and distance values are computed as for the regular delta plot. The values used for the item-total correlations for the form pilot and final form were the point biserial correlations of the item with all other items on the test. For the item pilot, the values were the correlations of the item with the original test items and the other pilot items appearing on the same form of the test. To determine the extent to which the original test and the revised test are related, students who were administered both the item pilot and the form pilot were selected and their test scores correlated. Correlations were .63 (N=102) for reading, .57 (N=102) for mathematics, and .41 (N=102) for language.

Stricker's Partial Correlation Index: Stricker's partial correlation index (Stricker, 1982) provides a measure of the association between group membership and item performance. Item responses and group membership are coded 1 or 0, and these two variables are correlated. Two additional correlations are then calculated. The first is a corrected correlation of the item response with the total score, and the second is the correlation of subgroup standing with the total score. Both correlations are corrected for attenuation. In this study the correction for attenuation was computed using the formula given by Ferguson (1981). The partial correlation index is then computed using the formula given by Stricker (1981).

Direct Comparison of Item Difficulties: The simplest and most straightforward of the six methods compared is a direct comparison of p values. In this procedure, item p values are computed separately for each subgroup. The minor subgroup p value is subtracted from the major subgroup p value, and the resulting p value difference is examined. Such a value is the basis for decisions regarding an item when the process associated with the Golden Rule procedure is applied.

Mantel-Haenszel Procedure: The Mantel-Haenszel procedure, originally used by Mantel and Haenszel (1959) in the biostatistics arena, is a chi-square technique that was suggested recently for use in education (Holland & Thayer, 1986; McPeek & Wild, 1986). Examinees in each subgroup are first divided into groups based on ability as estimated by test score. Recommendations concerning the number of ability groups to use for the Mantel-Haenszel procedure are not widely available, but Scheuneman's chi-square procedure includes a recommendation for three to five groups. Three groups were used in the present analysis because the majority of scores tended to fall within a small range at the upper end of the scale and because for the form pilot, the number of examinees was fairly small. The procedure calls for a 2 X 2 contingency table to be created for each item at each ability level. Cell values are the number of examinees in each subgroup at each ability level responding correctly or incorrectly to the item. Using the cell values for each of the ability levels, a Mantel-Haenszel chi-square statistic is computed using the formula, given by Holland and Thayer (1986).

Data

Data for this comparison study were the items from tests administered as part of Alabama's Basic Competency Education (BCE) program, which was first established in 1977. The first Basic Competency Tests (BCT) in reading, mathematics, and language were administered in Grades 3, 6, and 9 in the spring of 1980 (Teague, 1982, 1989). The current version of the Grade 9 BCT was first administered in the spring of 1990 following a spring 1989 item pilot and a fall 1989 form pilot.

Items from the spring 1989 item pilot, the fall 1989 form pilot, and the spring 1990 final form of the Grade 9 BCT reading, mathematics, and language tests were included in analyses completed to respond to the first two research questions. The third research question was addressed by examining items

common to all three administrations. For the spring 1989 item pilot, items being piloted were given as supplemental items during the regular test. The number of items appearing on each of 36 different supplemental forms (12 for reading, 13 for mathematics, and 11 for language) ranged from 18 to 25 new items. The fall 1989 form pilot consisted of 96 reading, 109 mathematics, and 102 language items. The final form in the spring of 1990 included 80 reading, 100 mathematics, and 96 language items. There were 72 reading, 89 mathematics, and 90 language items common to all three administrations.

Approximately 50,000 to 55,000 students participate in the Grade 9 BCT statewide during the spring of each year. For the spring 1989 item pilot administration, approximately 1,500 students responded to each pilot item. At least one form containing the supplemental items was distributed to each of the 132 Alabama school systems based on a stratified random assignment procedure (D. J. Steele, personal communication, February 1, 1991).

The fall 1989 form pilot was given to approximately 500 students. One system was selected randomly from each of eight clusters. Within each system, a school was selected randomly from the list of available schools within that system. The clusters from which systems were selected were created based on prior analyses that classified systems based on enrollment and three economic indicators (R. E. Lockwood, personal communication, February 19, 1991).

Six bias indices were computed for each of the spring 1989 pilot items and all items appearing on the fall 1989 form pilot and the spring 1990 administrations of the Grade 9 BCT for each of four subgroup pairs. The number of examinees included in the analyses for the item pilot ranged from 1,014 to 1,436 for each of the 36 forms. Males and females were about equally represented. White examinees outnumbered black examinees by about two to one. Examinees included in the fall 1989 form pilot analyses numbered 474 for reading, 468 for mathematics, and 466 for language. The spring 1990 administration included a total of 46,054 students for reading, 46,117 for mathematics, and 45,919 for language.

Procedures for Comparing Bias Indices

Addressing the first research question involved the selection of 20% of items with the highest bias index for each method at each test, administration, and subgroup pair. Selecting items at the 20% point

sometimes meant selecting in the middle of a set of tied ranks for bias indices. This was especially true of the partial correlation index and the direct difficulty differences. Thus, it is possible that an item was excluded whose bias index and rank were identical to that of a selected item. This situation was unavoidable because an examination of the ranks revealed that there was rarely a reasonable cut point (i.e., a point at which sufficient but not excessive numbers of items were selected) that did not involve ties in at least one set of ranks. The 20% cutoff resulted in either 4 or 5 items being selected for each of the item pilot forms and from 16 to 20 items being selected for each of the form pilot and final form tests.

The second research question was addressed by completing two types of computations. First the bias indices were ranked, and Spearman rank-order correlations were computed between the pairs of ranks for each of the tests, administrations and subgroup pairs. Percentage agreement in selection of items was also assessed by computing the percentage of biased items (as identified for the first research question) identified by both methods for each pair of methods for each test, administration, and subgroup pair.

The third research question was answered using similar procedures. First, bias indices for items common to all three administrations were extracted from the larger groups of indices for all items. These bias indices were then ranked for each technique and for each test, administration, and subgroup pair. Spearman rank-order correlations among the pairs of ranks for each administration pair were then computed. Next, the 20% of items with the highest bias index among the common items were identified. This amounted to 15 reading, 18 mathematics, and 18 language items. Finally, percentage agreement in item identification was computed between pairs of administrations for each test, method, and subgroup pair.

Results

Comparisons Among Techniques

Correlations: Values varied widely among the 168 correlations computed for each pair of techniques, representing almost the full range of possible values from -.76 to 1.00. Examination of the empirical frequency distribution of correlations revealed a large number of correlations at or above .60 with a drop in frequencies between .50 and .60. Therefore, .60 was used as the cutoff in reporting results.

Closer examination of the correlations revealed many consistent patterns. The strongest correlations were found between distractor analysis and direct difficulty differences. Of the 168 correlations, 156 were .60 or above, and 46 of those were at or above .90. A similar relationship emerged between the partial correlation index and the Mantel-Haenszel procedure, although the correlations were not as consistently high. A total of 109 correlations equaled or exceeded .60. Among the remaining pairs of methods, the four that showed the most consistent relationships were the partial correlation index with direct difficulty differences (94 correlations equal to or greater than .60), the partial correlation index with distractor analysis (83 correlations equal to or greater than .60), direct difficulty differences with the Mantel-Haenszel technique (78 correlations equal to or greater than .60), and distractor analysis with the Mantel-Haenszel technique (77 correlations equal to or greater than .60). In each of these comparisons, the majority of correlations equaling or exceeding .60 were evident among the female-male comparisons rather than among the black-white comparisons. The correlations between the Angoff delta plot technique and the Mantel-Haenszel technique, the partial correlation index, direct difficulty differences, and distractor analysis resulted in 51, 30, 24, and 11 correlations equal to or greater than .60 respectively. Fewer such correlations were found in the black-white comparisons than in the female-male comparisons. The lowest relationships were found between the modified Angoff technique and each of the other five techniques. None of the comparisons resulted in more than eight correlations equal to or greater than .60. The modified Angoff technique was also the technique that resulted in a very large number of negative correlations with other techniques.

Identification of the Highest 20% of Items: Examination of these data revealed that similarities exist among the techniques in their identification of items. Commonalities among the methods are more apparent among the most extreme items selected for the fall 1989 form pilot and the spring 1990 final form. For example, on the spring 1990 administration, two reading items were identified by all six techniques as being among the most extreme items for the two black-white comparisons. An example of a less obvious commonality was the identification of a reading item by all six techniques in the comparison between black females and males and by five out of six techniques in the comparison between white

females and males. Examination of extreme items for the spring 1989 item pilot involved fewer items per form and group. In these instances, slight differences in item position meant less of an opportunity to detect similarities among techniques in the specific items identified when only the highest 20% of items were selected. Still, repeated identification of a given item could be detected. For example, one reading item was identified by five out of six methods in both of the black-white comparisons, by four out of six methods for the black female-male comparisons, and by six out of six methods for the white female-male comparisons. Examination of the administration history of the items revealed that this item was not selected for use following the item pilot.

Percentage Agreement: The results of these analyses were consistent with the findings based on correlations among the ranks of the bias indices. The greatest agreement across all groups, tests, and administrations was found between distractor analysis and direct difficulty differences. In this comparison 138 of the 168 percentage values were greater than 50%. The comparison between the partial correlation index and the Mantel-Haenszel technique resulted in 121 values greater than 50%.

Comparisons of direct difficulty differences with the partial correlation index, distractor analysis with the partial correlation index, direct difficulty differences with the Mantel-Haenszel technique, distractor analysis with the Mantel-Haenszel technique, and the Angoff delta plot with the Mantel-Haenszel technique yielded 102, 106, 95, 110, and 89 values above 50% respectively. As with the correlations for these pairs, the percentage agreement was generally higher for the female-male comparisons than for the black-white comparisons.

Only moderate agreement was found between the Angoff delta plot and three other techniques. Percentage agreements calculated between the Angoff delta plot and the partial correlation index, distractor analysis, and direct difficulty differences resulted in better than 50% agreement 55, 46, and 42 times respectively. Lowest agreement was found in the comparisons of the modified Angoff with all others. The number of values greater than 50% ranged from 2 to 18. Interestingly, the percentage agreement between the Angoff and modified Angoff was 100% on two occasions, an inconsistent finding in the context of the other comparisons of the modified Angoff with all other techniques.

Comparisons Across Administrations

Correlations: Again, consistent patterns were apparent. The strongest correlations between pairs of administrations were found for direct difficulty differences. Nearly half of the 36 correlations for this technique were at or above .60. Another 12 were from .30 to .60. Most of the higher correlations were found among the black-white comparisons. A similar pattern was found among the correlations for distractor analysis. Eleven were greater than .60, and eight of those were among the black-white comparisons. Another 16 correlations for distractor analysis were between .30 and .60. Although not as strong or consistent, four correlations for the partial correlation index were above .60, and another 14 were between .30 and .60. The Mantel-Haenszel technique had no correlations above .60, but 15 correlations fell between .30 and .60. The lowest correlations were found for the modified Angoff technique. Only two correlations were between .30 and .60, and 7 of the 24 correlations were negative. Angoff delta plot analyses resulted in only a few more correlations between .30 and .60, and no correlations greater than .60.

Identification of the Highest 20% of Items: As with the comparisons among techniques, similarities were observed across administrations for each of the techniques. For example, four of the techniques identified one reading item three times for the black-white female comparisons. Three techniques identified the same item all three times for the black-white male comparisons. A similar pattern was evident for another reading item for the male-female comparisons.

Percentage Agreement: Greater agreement was evident between pairs of administrations for distractor analysis and for direct difficulty differences than for any of the other techniques. Generally the agreements were higher for black-white comparisons than for female-male comparisons. The lowest percentage agreement was found for both Angoff techniques. These results are consistent with outcomes of the correlation analyses.

Discussion

This study represents one of the few efforts to compare item bias techniques in the context of criterion-referenced minimum competency tests. Only three other such studies were found. An early study by Nungester (1977) compared Angoff's delta plot to two other methods, Perlman et al. (1988) examined

the Mantel-Haenszel technique and Angoff's delta plot among others, and Hambleton and Rogers (1989) included the Mantel-Haenszel technique in their study. The four other techniques reported in this study have not appeared in previous research on criterion-referenced minimum competency tests. None of the research reported in the literature has addressed the stability of the techniques from item tryout to final form on criterion-referenced minimum competency tests.

Discussion of Results for Each Technique

Distractor Analysis: Results for distractor analysis indicated that its strongest relationships were with direct difficulty differences, the partial correlation index, and the Mantel-Haenszel procedure. The relationship with direct difficulty differences should not be surprising because both procedures examine group differences in responding to an item without considering differences in ability levels for the two groups. The direct difficulty difference method does so through examination of p value differences, while distractor analysis involves a chi-square statistic to determine the response choice with the greatest difference between groups in response rate. If a large p value difference exists between groups, the likelihood of an item's being identified by the distractor analysis technique is greater.

The relationship between distractor analysis and the Mantel-Haenszel procedure was somewhat more evident for the female-male comparisons. This pattern was even more pronounced for the relationship between distractor analysis and the partial correlation index. Although not immediately apparent from the data reported here, one possible explanation is that the score differences of black and white examinees tend to be slightly greater than those of male and female examinees. Since both the Mantel-Haenszel technique and the partial correlation index control for ability, it would seem reasonable to expect that these procedures would yield results similar to unconditional procedures such as distractor analysis when group differences in ability are smaller.

Angoff Delta Plot: The strongest relationship of this method with another was with the Mantel-Haenszel procedure, although it can be considered moderate at best. The delta plot method does take into account group ability differences by using the major axis as the point from which distances are computed. However, the results of the present study support the notion that the Angoff delta plot

procedure and the Mantel-Haenszel technique identify slightly different sets of items based on different definitions of item bias. The Angoff procedure is designed to identify items with relatively greater differences than other items in a given set. The Mantel-Haenszel technique is designed to consider each item independently.

Present findings are not inconsistent with those of Perlman et al., (1988) who found correlations of .75 to .92 between the Angoff delta plot and the Mantel-Haenszel technique. However, many correlations in the present study were quite low, indicating that the relationship between the two methods is not as consistent as would be suggested from the results of Perlman and her colleagues.

Modified Angoff Delta Plot: Results of the present study indicated little or no relationship between this method and any other method examined. In fact, many of the correlations were negative, indicating that this method was selecting a different set of items than any of the other methods. On two occasions, however, perfect correlations were found between this method and the Angoff delta plot. Such results indicate substantial inconsistency in the method.

A post hoc inspection was conducted for a sample of the items identified by the modified Angoff method. The general observation was that the items being selected were very difficult items, often with relatively low discrimination indices, but not necessarily with differing difficulty levels for subgroups. In a few instances, the items selected were very easy items with no evidence of differential functioning among groups. For example, one item had a p value of .95 with percentage responses for the three distractors of 2%, 2%, and 1%. These were not the types of items one would expect to identify as biased against one group or another.

Present findings support those of Seong and Subkoviak (1987) who found that the modified delta plot correlated only .35 with an a priori index of item bias. They also found that the modified Angoff demonstrated lower agreement than other methods studied in the selection of 10 biased items. Shepard et al. (1985) also indicated the unsatisfactory performance of the modified Angoff, but they did not report the nature of their findings.

Partial Correlation Index: The partial correlation index was most highly related to the Mantel-Haenszel technique. This method also resulted in fairly strong relationships with distractor analysis and direct difficulty differences. The relationship with the Mantel-Haenszel technique tended to be only slightly greater for female-male comparisons than for black-white comparisons, but the other two relationships were more evident among the female-male comparisons. As was discussed earlier, the greater number of higher correlations for the female-male comparisons may very well be due to smaller group ability differences.

In the present study, the agreement between the partial correlation index and the Angoff delta plot was rather weak. This finding supports the results Stricker (1982) obtained.

A limitation of this method may be revealed in a comment made by Stricker (1982). He indicated that an assumption of the method is the "lack of differential functioning in the total score" (p. 262). This assumption seems inconsistent with the aim of the method which, as Stricker stated, is to "[control] for subgroup differences in overall ability" (p. 262). The requirement that such an assumption must be met may, however, explain why the partial correlation index compared somewhat more favorably with the Mantel-Haenszel technique among the female-male comparisons where the ability differences were presumably smaller.

Direct Difficulty Difference: The strongest relationships involving this procedure were with distractor analysis and the partial correlation index, as discussed above, and the Mantel-Haenszel technique. Among the correlations with the Mantel-Haenszel technique, the greatest number of higher correlations occurred among the female-male comparisons. Again, smaller group ability differences between females and males might explain this finding. It is widely accepted that, as with distractor analysis, lack of control for group ability differences is a limitation of this procedure.

Mantel-Haenszel Procedure: Of all the methods studied, the one that showed fairly consistent relationships with the greatest number of other techniques was the Mantel-Haenszel procedure. This method correlated .60 or greater over half the time with all other procedures except the modified Angoff. There was a tendency across all methods for the relationships to be equal to or greater than .60 more

frequently among the female-male comparisons. The tendency was more pronounced in the relationships between the Mantel-Haenszel and the unconditional procedures such as distractor analysis and direct difficulty differences, again possibly due to the smaller group ability differences.

Discussion of Comparisons Across Administrations

The method demonstrating the greatest stability across administrations was the direct difficulty difference, followed closely by distractor analysis. The Mantel-Haenszel procedure and the partial correlation index both revealed weak to moderate stability. The poorest results with regard to stability were found for both Angoff techniques.

Present results generally support previous findings, even when the cross-sample comparisons are made under somewhat different conditions. Hoover and Kolen (1984) found the Angoff delta plot method to be quite unreliable across samples, yielding values of .07 and .29. Stricker (1984) found low agreement of the Angoff delta plot from sample to sample and moderate stability for the partial correlation index. Perlman et al. (1988) demonstrated that the Angoff delta plot's reliability ranged from low to moderate depending on sample size.

Two groups of researchers examining the reliability of the Mantel-Haenszel technique achieved results similar to those of the present study. Perlman et al. (1988) reported low to moderate reliabilities for the Mantel-Haenszel, again depending on sample size. Hambleton and Rogers (1989) reported instability of the Mantel-Haenszel technique across two diverse cultural groups.

Sample size may indeed have been a variable influencing stability in the present study in the case of the four techniques with low agreement across administrations. Certainly the work of Perlman and her associates (1988) demonstrated that for cell sizes of 1,000, reliabilities were much better than for cell sizes of 200. Stricker (1984) also recommended limiting the use of the partial correlation index to samples of at least 1,500 and cell sizes of 300 at least until the properties of the measure are better understood. The present analyses typically involved cell sizes of under 400 for two of the three administrations of items, and some cell sizes were under 200.

Another likely explanation for the low stability of the Angoff techniques especially, is the dependence of these techniques on the context in which the items are examined. Particularly for the spring 1989 item pilot, the context for the item bias computations was very different from that of the later administrations. Items were compared to approximately 17 to 24 other items that were very closely related in content and form. Later administrations involved a larger number and a more diverse set of items, and although the fall 1989 form pilot and the spring 1990 final form were very similar, they were not identical.

Still another possible reason for low agreement between administrations, especially for the Mantel-Haenszel technique and the partial correlation index, relates to the total score used in computations of the indices. For the spring 1989 item pilot, total score was computed for items on the old BCT. The fall 1989 form pilot and the spring 1990 final form items were used to compute total score for those administrations. As was stated earlier, the correlations between the old test and the form pilot ranged from .41 to .63.

Conclusions

The clearest conclusion that can be drawn from the results of this study is that the modified Angoff method does not appear to be a desirable technique. The sets of items identified by the modified Angoff technique were clearly different from those identified by any of the other techniques. The nature of the items identified, based on observations discussed earlier, suggested that it was the modified Angoff method and not the other methods that identified an inappropriate set of items.

Results support the possible use of more than one index to examine a set of items. The most stable techniques were those that should not be used by themselves. As Ironson et al. (1984) indicated, direct difficulty difference is not accepted in item bias research as an appropriate procedure to use alone because of the absence of control for ability differences. The method did, however, bear some relationship to other methods, and it provided the most stable measure across administrations among those methods included in this study. Distractor analysis was very closely related to direct difficulty difference in this study and may be useful as an alternative to direct difficulty difference or, as was suggested by Burrill

(1981), as a follow-up to any item bias procedure. In fact, distractor analysis may be preferable because of the additional information it can provide.

Stricker's stated assumption for using the partial correlation index (lack of differential total test score functioning), as well as the limitations he suggested with regard to sample size, pose limitations to the use of this procedure. The partial correlation index appears to warrant further investigation.

The Mantel-Haenszel technique appears to be an appealing choice. It was related to other methods in the study, but not so highly related that it identified the same items. Although it was not highly stable across administrations, it was somewhat more stable than the Angoff delta plot method.

Recommendations for Further Study

1. Several investigators in the past have suggested that signed indices result in higher correlations among methods. Signed indices may possibly result in higher agreement across administrations. The present study examined only the unsigned indices. Future research comparing present results to outcomes using signed indices would determine if higher correlations are obtained for items used in this study.

2. Special education students were excluded from present analyses. Analysis that includes special education students would reveal the extent to which their inclusion would have affected outcomes.

3. The present study included comparisons among race and sex subgroups as they existed in the population. Results suggested that ability differences may be a contributing factor in the outcomes of the present study. Similar analyses for these data using matched groups might help to address the question of whether some of the patterns of correlations were, in fact, due to group ability differences.

4. Sample size appears to play a role in the outcomes. More studies, like that of Perlman and her colleagues, and including some of the techniques used in the present study, would provide additional information concerning the effects of sample size.

5. An issue with the Mantel-Haenszel technique involves what total score to use in creating the ability levels. Holland and Thayer (1986) indicated that some preliminary work was being done to investigate the advisability of excluding from the matching criterion items that, on preliminary analysis, exhibit differential functioning. No such studies were found in the literature.

References

Anastasi, A. (1982). Psychological testing (5th ed.). New York: MacMillan Publishing Co., Inc.

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

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 Press.

Angoff, W. H. (1988). Proposals for theoretical and applied development in measurement. Applied Measurement in Education, 1, 215-222.

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

Angoff, W. H., & Sharon, A. T. (1974). The evaluation of differences in test performance of two or more groups. Educational and Psychological Measurement, 34, 807-816.

Berk, R. A. (1984). Conducting the item analysis. In R. A. Berk, (Ed.), A guide to criterion-referenced test construction (pp. 97-143). Baltimore: Johns Hopkins University Press.

Burrill, L. E. (1981). A comparative investigation into the identification of ethnic bias in items assessing current educational status. (Doctoral dissertation, Fordham University, 1981). Dissertation Abstracts International, 42, 1110A.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rinehart and Winston.

Denton, L. (1988). Board votes to oppose Golden Rule technique. APA Monitor, 19(5), 7.

Devine, P. J., & Raju, N. S. (1982). Extent of overlap among four item bias methods. Educational and Psychological Measurement, 42, 1049-1066.

Eells, K., Davis, A., Havighurst, R. J., Herrick, V. E., & Tyler, R. W. (1951). Intelligence and cultural differences: A study of cultural learning and problem-solving. Chicago: The University of Chicago Press.

Faggen, J. (1987). Golden Rule revisited: Introduction. Educational Measurement: Issues and Practice, 6(2), 5-8.

Ferguson, G. A. (1981). Statistical Analysis in Psychology and Education (5th ed.). New York: McGraw-Hill.

Hambleton, R. K., & Rogers, H. J. (1989). Detecting potentially biased test items: Comparison of IRT area and Mantel-Haenszel methods. Applied Measurement in Education, 2, 313-334.

Holland, P. W., & Thayer, D. T. (1986). Differential item functioning and the Mantel-Haenszel procedure. Program Statistics Research Technical Report No. 86-69. Princeton, NJ: Educational Testing Service.

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

Intasuwan, P. (1979). A comparison of three approaches for determining item bias in cross-national testing (Doctoral dissertation, University of Pittsburgh, 1979). Dissertation Abstracts International, 40, 2613A-2614A.

I.Q. tests banned in California. (1987, Spring). Fairtest Examiner, 1(1), 10.

Ironson, G. H. (1982). 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 Press.

Ironson, G. H., & Craig, R. (1982). Item bias techniques when amount of bias is varied and score differences between groups are present. Final Report. NIE Report G-81-0045. (ERIC Document Reproduction Service No. ED 227 146)

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

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

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

Jensen, A. R. (1984). Test bias: Concepts and criticisms. In C. R. Reynolds and R. T. Brown (Eds.), Perspectives on bias in mental testing (pp. 507-586). New York: Plenum Press.

Laksana, S. (1979). Application of analysis of variance approach and item characteristic curve approach for assessing item bias in the ITBS Form 7. (Doctoral dissertation, University of Iowa, 1979). Dissertation Abstracts International, 40, 2615A-2616A.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute, 22, 719-748.

McPeek, W. M., & Wild, C. L. (1986, April). Performance of the Mantel-Haenszel statistic in a variety of situations. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Mellenbergh, G. J. (1981). Conditional item bias methods. In S. H. Irvine & J. W. Berry (Eds.), Human assessment and cultural factors (pp. 293-302). New York: Plenum Press.

Merz, W. R., & Grossen, N. E. (1978). An empirical investigation of six methods for examining test item bias. Final Report. NIE Report 6-78-0067. (ERIC Document Reproduction Service No. ED 178 566)

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

Osterlind, S. J. (1983). Test item bias. SAGE University paper series: Quantitative Applications in the Social Sciences. 30.

Overcast, T. D., & Sales, B. D. (1982). The legal rights of students in the elementary and secondary public schools. In C. R. Reynolds & T. B. Gutkin (Eds.), The handbook of school psychology (pp. 1075-1100). New York: John Wiley & Sons.

Perlman, C. L., Bezruczko, N., Junker, L. K., Reynolds, A. J., Rice, W. K., & Schulz, E. M. (1988, April). Investigating the stability of four methods for estimating item bias. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, Louisiana.

Peterson, N. S. (1980). Bias in the selection rule--bias in the test. In L. J. T. van der Kamp, W. F. Langerak, & D. N. M. de Gruijter (Eds.), Psychometrics for educational debates (pp. 103-122). New York: John Wiley & Sons.

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.

Rebell, M. A. (1988). Legal issues concerning bias in testing. In R. G. Allan, P. M. Nassif, & S. M. Elliot (Eds.), Bias issues in teacher certification testing (pp. 1-18). Hillsdale, NJ: Lawrence Erlbaum Associates.

Reynolds, C. R. (1982a). Methods for detecting construct and predictive bias. In R. A. Berk (Ed.), Handbook of methods for detecting test bias (pp. 199-227). Baltimore: Johns Hopkins University Press.

Reynolds, C. R. (1982b). The problem of bias in psychological assessment. In C. R. Reynolds & T. B. Gutkin (Eds.), The handbook of school psychology (pp. 178-208). New York: John Wiley & Sons.

Rooney, J. P. (1987). Golden Rule on "Golden Rule". Educational Measurement: Issues and Practice, 6(2), 9-12.

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

Rudner, L. M. (1978a). An evaluation of select approaches for biased item identification. (Doctoral dissertation, The Catholic University of America, 1977). Dissertation Abstracts International, 38, 5410A.

Rudner, L. M. (1978b). Using standard tests with the hearing impaired: The problem of item bias. The Volta Review, 80, 31-40.

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

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

Scheuneman, J. D. (1981). A new look at bias in aptitude tests. In P. Merrifield (Ed.), Measuring Human Abilities: New Directions for Testing and Measurement, 12, 3-35.

Seong, T., & Subkoviak, M. J. (1987, April). A comparative study of recently proposed item bias detection methods. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, DC. (ERIC Document Reproduction Service No. ED 281 883)

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

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

Standards for Educational and Psychological Testing. (1985). Washington, DC: American Psychological Association.

Stricker, L. J. (1981). A new index of differential subgroup performance: Application to the GRE Aptitude Test. GRE Board Professional Report GREB No. 78-7P.

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

Stricker, L. J. (1984). The stability of a partial correlation index for identifying items that perform differentially in subgroups. Educational and Psychological Measurement, 44, 831-837.

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.

Teacher testing: Victory in Alabama. (1987, Spring). Fairtest Examiner, 1(1), 5.

Teague, W. (1982). Minimum standards and competencies (reading, language, and mathematics) for Alabama Schools (1982 edition). Bulletin 1982, No. 25. Montgomery, AL: Alabama State Department of Education.

Teague, W. (1989). Minimum standards and competencies (reading, language, and mathematics) for Alabama Schools (1989 edition). Bulletin 1989, No. 37. Montgomery, AL: Alabama State Department of Education.

Veale, J. R., & Foreman, D. I. (1983). Assessing cultural bias using foil response data: Cultural variation. Journal of Educational Measurement, 20, 249-258.

Welch, C. J., Ackerman, T. A., Doolittle, A. E., & Hurley, J. (1987, April). An examination of statistical procedures for detecting cross-cultural differential item performance. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, DC.

Wise, L. L. (1987, April). Differential item difficulty indicators in small samples. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.