DOCUMENT RESUME

ED 148 899 TM 006 870

AUTHOR Rudner, Lawrence M.

TITLE Efforts Toward the Development of Unbiased Selection and Assessment Instruments.

PUB DATE 28 Jun 77

NOTE 36p.; Paper presented at the International Symposium on Educational Testing (Third, University of Leyden, The Netherlands, June 28, 1977)

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.

DESCRIPTORS Analysis of Variance; Criterion Referenced Tests; *Cultural Differences; Culture Free Tests; Factor Analysis; *Identification; Individual Differences; *Item Analysis; *Mathematical Models; Norms; Response Style (Tests); Standardized Tests; Statistical Analysis; Student Testing; *Test Bias; *Test Construction; Test Items; Test Reliability

IDENTIFIERS Chi Square; Distractors (Tests); Item Characteristic Curve Theory; Latent Trait Theory

ABSTRACT

Investigations into item bias provide an empirical basis for the identification and elimination of test items which appear to measure different traits across populations or cultural groups. The Psychometric rationales for six approaches to the identification of biased test items are reviewed: (1) Transformed item difficulties: within-group p-values are standardized and compared between groups. (2) Analysis of variance: bias is operationally defined in terms of significant item by group interaction effects. (3) Chi-square: individual items are investigated in terms of between group score level differences in expected and observed proportions of correct responses. (4) Item characteristic curve theory: differences in the probabilities of a correct response, given examinees of the same underlying ability and different culture groups, are evaluated. (5) Factor analytic: item bias is investigated in terms of culture specific and culture common sources of variance, or in terms of loadings on a biased test factor. (6) Distractor response analysis: the relative attractiveness of item foils, or response sets, is investigated. The limitations and advantages of each approach in terms of the underlying assumptions, psychometric soundness, conceptual complexity, applicability to criterion referenced tests and applicability to interdependent groups are discussed. (Author/MV)

EFFORTS TOWARD THE DEVELOPMENT OF UNBIASED SELECTION AND ASSESSMENT INSTRUMENTS

Lawrence M. Rudner

Gallaudet College
The Model Secondary School for the Deaf
Kendall Green
Washington, D.C. 20002

Paper presented at the Third International Symposium
on Educational Testing, University of Leyden, The Netherlands

June 28, 1977

ABSTRACT

Different methodologies have been proposed for the evaluation of bias both in selection and assessment instruments and in the items within such measures. While bias in an instrument as a whole is of prime concern to test users and has received considerable attention in popular and professional literature, bias in test items is of increasing concern to test developers. Investigations into item bias provide an empirical basis for the identification and elimination of items which appear to measure different traits across population/culture groups. Thus, they help to decrease bias in instruments under development.

This paper reviews the psychometric rationales of the following six types of approaches to biased item identification:

1. transformed item difficulties approaches in which within-group p-values are standardized and compared between groups.
2. analysis of variance approaches in which bias is operationally defined in terms of significant item by group interaction effects.
3. chi-square approaches in which individual items are investigated in terms of between group score level differences in expected and observed proportions of correct responses.
4. item characteristic curve theory approaches in which differences in the probabilities of a correct response, given examinees of the same underlying ability and different culture groups, are evaluated.
5. factor analytic approaches in which item bias is investigated in terms of culture specific and culture common sources of variance or in terms of loadings on a biased test factor.
6. distractor response analysis approaches in which the relative attractiveness of item foils is investigated.

Limitations and advantages of the approaches in terms of their underlying assumptions, psychometric soundness, conceptual complexity, applicability to criterion referenced tests and applicability to interdependent groups are discussed and evaluated.

Efforts Toward the Development of Unbiased Selection and Assessment Instruments*

Approximately 25 years ago Eells and his colleagues conducted what appears to be the first serious attempt to examine test items for bias (Eells, et.al., 1951) and developed one of the first measures purported to be culture fair. Since that time, the entire issue of cultural bias in measurement has become heated, complex, and pronounced in the literature. Actions by the National Association of Black Psychologists, the American Personnel and Guidance Association, the National Education Association, the National Association for the Advancement of Colored People, the National Association of Elementary School Principals and the Council of the Society for the Psychological Study of Social Issues calling for moritoria on certain types of tests, banning tests, and requiring alternate plans for testing, indicate the serious nature of the current situation (see Williams, Mosby and Hinsen, 1976). The concern is also apparent in recent litigation (DeFunis vs. Odegaard, 1974; Diana vs. the California State Board of Education, 1970; Hobsen vs. Hansen, 1967). Naturally, all this has not gone unnoticed by those involved in the measurement field. Bias and debiasing studies have occurred and various models been proposed in ever-expanding efforts to meet the challenge of bias in educational assessment.

One major type of bias investigations is concerned with the instrument as a whole and examines the question: Does a test unduly favor or impede examinees from different parts of the country or of different backgrounds? Another is concerned with the items within a test and asks: Which items and

*The author is indebted to David Knight and to William Merz for their valuable assistance with earlier drafts of this report, to Sonya Johnson and Eileen Roper for their editorial assistance and to Jacqueline Cox and Marianne Walker for typing this manuscript.

item formats are appropriate for a given population and which may be used across given cultures?

The first type of investigation is of interest to the test users who need to know the accuracy of the test information. The models proposed by Cleary (1968), Thorndike (1971), Darlington (1971), Cole (1973), Einhorn and Bass (1971) and Gross and Su (1975) (also see the entire spring 1976 issue of the Journal of Educational Measurement) exemplify this first type of investigation. The second type of investigation is of interest to developers as it assists them in developing valid and cross-culture fair items and provides a framework for constructing better tests in subsequent efforts. The work of Angoff (1972), Cardall and Coffman (1964), Green and Draper (1972), Merz (1973, 1976a), Rudner (1977a), Scheuneman (1975) and Veale and Foreman (1975, 1976) have been directed at this need. It is this second type of bias, item bias, which the present paper addresses.

Bias and the Item Tryout Procedure

Test and item bias generally stem from two major sources, the human element involved in test development and the procedures used to evaluate the test and test items. The first source of bias stems from cultural differences between test developers and some test users. That is, the cultural incongruity between test developers and users may subtly manifest itself in items which are insensitive to the experiences, morals, and thinking of particular cultural groups. Efforts by test developers to include members of various cultural groups in the development and review of items will help identify some biased items (see Green, 1971; Fitzgibbons, 1971), but certainly not all.

The second source of bias comes into play when data from a population sample are used to improve the effectiveness of a test (Green, 1972). This procedure, which as Green points out, has not changed in 50 years (Cf. Ruch, 1929, Chapter 2; Lord and Novick, 1974, Chapter 15) is basic to the development of effective achievement tests. However, during the item-tryout, the characteristics of the dominant group will tend to overshadow those of minority groups. As a result, items which are most sensitive to the abilities, cognitive styles, and knowledge of the dominant group are selected. Such items may be biased against the examinees whose attributes diverge from those of the collective item-tryout sample.

The development of a standardized measure typically involves the administration of a carefully developed item pool to a large representative sample of examinees whose attributes are similar to those of the intended population of examinees. Typically, a measure of each item's discrimination power, e.g., the item-test point biserial correlation, is computed and those items discriminating best are retained. As the population of this country is largely white middle class, the items most sensitive to white middle class attributes are those which are most often retained.

Green (1972) examined a few questions about this procedure: Are different items retained when different culture groups compose the item-tryout sample? Will scores differ using tests composed of uniquely retained items? Will test reliabilities differ using the different tryout samples?

Using the different levels and subtests of the California Achievement Test battery as item pools and different subgroups of the standardization sample as item-tryout samples, Green computed separate sets of item-test point biserial corrections.[1] From each set of correlations, the best half of the items (those with the highest correlations) were noted and pairwise comparisons made. Aberrant items were then defined as those items retained based on one subgroup of a pair, and rejected based on the other.

If all the subgroups responded to the items in the same manner, identical items would be retained. However, this did not occur. The overall median proportion of identical items which were retained in comparing all 21 possible pairs of item-tryout samples was only .70--a relatively low percentage. Clearly, different item-tryout samples from different cultural backgrounds lead to the selection of different items.

This, in itself, is not disturbing. Since the point biserial correlation is partly a function of item difficulty, one might expect a number of items to be uniquely related. However, suppose different items are retained for whites and blacks, and blacks obtain dissimilar total scores using (1) the items uniquely retained based on blacks and (2) the items uniquely retained based

[1] CTB/McGraw Hill (1974) and Ozenne, Van Gelder and Cohen (1974) have used discordant point biserial correlations as a method of identifying biased items in developing and restandardizing national achievement tests.

on whites. This would be cause for alarm. In comparing such sets of scores, Green found correlations ranging from -.17 to +.82 with a median of about .5. Since the number of items in these tests composed of uniquely retained items were less than the original item pool, the reliabilities were corresponding low. However, even after correcting for attenuation (bringing the median correlation to about .8), large amounts of variance in each set were still unaccounted for. These different scores indicate that the unique items taken collectively may measure ability differently across populations.

Green also computed the Kuder-Richardson reliabilities (KR-20) of the item pools using different cultural groups. Differing reliabilities would indicate that the scores of one cultural group contain more error than those of another cultural group. The median KR-20 reliabilities in Green's study were all .92± .02. Clearly, there was little evidence of bias by this criterion. Perhaps this was because measures of internal consistency, such as the KR-20, are largely sensitive to test length (Guilford, 1954, pp. 352-353).

In summation, Green showed that different items most probably will be selected when different cultural groups are used as the item-tryout sample and that scores obtained from these uniquely selected items will differ, even though the item pools exhibit high degrees of internal consistency. The task, then, is to modify the test development procedure so that items which are unduly sensitive to cultural differences can be identified and either revised or eliminated.

Approaches to Biased Item Identification

Recently, procedures and models have been proposed and advocated for identifying biased items within a test: (1) analysis of variance approaches, (2) transformed item difficulties (p-values) approaches, (3) chi-square approaches, (4) item characteristic curve theory approaches, (5) factor analytic approaches, and (6) distractor response analysis approaches. The interested reader is referred to Green and Draper (1972) for an empirical investigation using and comparing a few of the earlier approaches within the second through fifth categories and to Merz (in preparation) and Rudner (in preparation) for empirical investigations comparing some newer approaches.

Analysis of Variance Approaches

In the first type of approach, which defines bias as a significant item by group interaction, subjects sampled from two or more populations are given a common test and the resultant variations in item scores are analyzed by an analysis of variance design. Variance could be attributed to differences in (1) items, as some items are more difficult than others; (2) groups, as one group may have more of the measured attribute than another; (3) subjects within groups, as examinees will differ in ability; and (4) an interaction of the items and the groups. When the groups are defined by cultural affiliations, a significant item by culture interaction is indicative of some items being relatively more difficult for members of one culture than another. Post hoc testing procedures, such as Duncan's Multiple Range Test (Duncan, 1955, 1957), can be used to identify specific items showing bias in terms of significant differences in relative item difficulty.

Examples of this approach are found in Cardall and Coffman (1964), Cleary and Hilton (1968), Eagle and Harris (1969), Hoeptner and Strickland (1972), and Jensen (1973). In order to use this approach properly, extremely large sample sizes are required in order to control for variables such as IQ, socio-economic status, parental education level, ethnicity, and attitudes. However, this is true for all investigations into item and test bias.

Jensen (1973) reported two studies in which he attempted control by matching subjects from different cultures on their mental age. In both studies great reductions were found in the item by culture interaction after matching, indicating that the procedure may be more sensitive to ability than to cultural variations.

Jensen confirmed this in a second investigation. After using an analysis of variance approach with white and black subjects (without matching mental age), he conducted a second analysis of variance using two groups of Caucasians whose score distributions closely matched those of the blacks and whites in the first part of the study. The results of this pseudo-race comparison closely matched those of the true race comparison, especially with regard to the item by culture interaction. He concluded that "it would be extremely difficult to make a case that the race by items interaction is attributable to cultural bias" (p. 17). Thus; Jensen claims that this procedure may be sensitive to differences in ability rather than to cultural differences.

Whether or not Jensen's claim is valid, two additional major problems with this approach exist. First, the practical alpha level in the post hoc analysis can become inflated as the number of items increases. Hence, one must be aware that some items may be erroneously classified as biased. The second and more serious problem arises from the underlying assumption that the total scores are unbiased. Inasmuch as the identification of biased items may contradict this assumption, the procedure poses some conceptual problems.

Transformed Item Difficulties Approaches

The transformed item difficulities approaches, providing for a visual examination of item by group interaction effects, were probably first described by Thurstone (1925) in connection with his method of absolute scaling. Of the approaches, this method appears to be one of the best known. It has been advocated and used frequently by Angoff, (1972; and Ford, 1974; and Modu, 1973) and others (Green and Draper, 1972; Jensen, 1973; Hicks, et al., 1976;

Strassberg-Rosenberg and Donlon, 1975; Echternacht, 1975; Rudner, 1977b). Further, the approach has appeared in at least one meassurement textbook (Anastasi, 1976, pp. 222-226).

> In this method, indices of item difficulty--i.e., p-values--are obtained for two different groups on a number of items. Each p-value is converted to a normal deviate and the pairs of normal deviates, one pair for each item, are plotted on a bivariate graph, each pair represented by a point on the graph (Angoff, 1972, p. 1).

The plot will generally be in the form of an ellipse. A 45 degree line, passing through the origin, provides a theoretical regression indicating the absence of bias. Items greatly deviating from this line may be regarded as exhibiting an item by group interaction. That is, relative to the other items, deviant items are especially more difficult for members of one group than the other. Assuming both groups received similar instructions, such items would appear to represent different psychological meanings for the two groups of examinees.

Since the intent is to make comparisons of between-group differences in item difficulty, it is necessary to transform the proportion passing an item to an index of item difficulty which constitutes at least an interval scale. This is accomplished by expressing each item p-value in terms of within-group deviations of a normal curve (see Guilford, 1954, pp. 418-419). Any linear transformation of the item z-score will meet such a requirement. One such transformation has been Delta values (4z + 13).

The distance of an item point to the line can be treated as a measure of the degree of item bias. One can determined which items are "greatly deviating" from the line by incorporating any of the traditional or nontraditional methods of outlier or residual analysis. One method is to place confidence limits on the line by using a multiple of the standard error of estimation. An alternate approach, adopted by Strassberg-Rosenberg and Donlon (1975) and Hicks, et al. (1976) involves computing the standard deviation of the residuals and classifying as biased those items deviating by greater than 1.5 standard deviation units. Rudner (1977b) has employed a fixed item-regression line distance of .75 z-score units.

An example of the approach is shown in Figure I. The transformed p-values have a correlation of approximately .90, making the plot relatively long and flat. The solid line represents the main axis and the dotted lines represent linear confidence limits. The item represented in the upper left, outside the confidence interval, would be considered biased.

As a modification of this procedure, Green and Draper (1972, p. 16) suggest that the "item-test biserial correlations might be incorporated . . . so as to estimate the linear test score-item score regression whereby item difficulties may be formed in a manner analogous to the way in which adjusted means are formed in an Analysis of Covariance." Since by this procedure, differential item discrimination indices and item difficulties would both influence item locations on the regression plot, items which have proportional p-values but disproportionate discrimination indices would have a greater tendency to deviate from the main axis of the scatterplot and show up as aberrant.

Chi-Square Approaches

A third approach to biased item analysis determines whether examinees of the same ability level have the same probability of a correct response regardless of cultural affilation. This is accomplished by dividing the tryout samples into groups based on their observed score and comparing the proportions of students within each level responding correctly with a chi-square test for independent observations (Scheuneman, 1975, 1976; Green and Draper, 1972). An item is considered unbiased if, for all individuals in the same total score

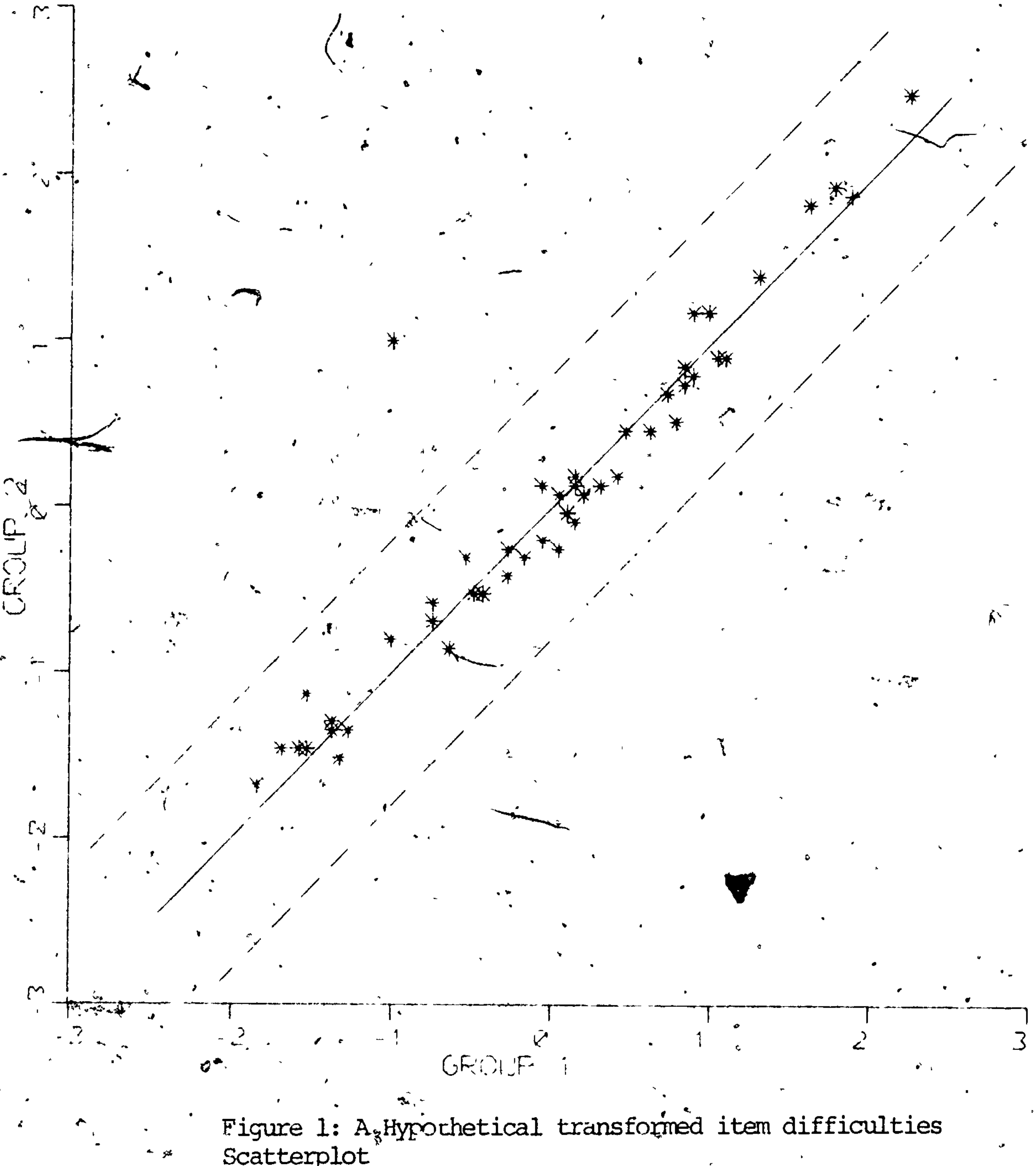

Figure 1: A Hypothetical transformed item difficulties Scatterplot

interval, the proportion of correct response is the same for both groups under consideration. A modified chi-square test determines the probability that an item is unbiased by this definition.

Scheineman (1976), in applying the approach to several sets of data, advocates using four or five total score levels based on the score distribution of the smaller sample (Green and Draper had used within-group quintiles). As with the analysis of variance approach, the procedure requires a large number of inference tests. Again, unbiased items may be misclassified as biased because of inflated alpha levels. Further, the procedure assumes total scores to be valid measures of ability and appears to be unduly sensitive to differences in the total score distributions of the examined samples.

Item Characteristic Curve Theory Approaches

Latent trait or item characteristic curve theory relates the probability of a correct item response to a function of an examinee's underlying ability level (θ_i) and characteristic(s) of the item. While the various models (Lord, 1952; Rasch, 1960; Birnbaum, 1968; Urry, 1970) differ in terms of the number of item parameters considered; they all describe the item parameter(s) independently of the examined sample. This attractive property has led to the development of some interesting applications in test development, adaptive testing and equating, and may prove useful in detecting item bias.

One general, cumulative logistic model formalized by Birnbaum uses three item parameters: a_g - an item discrimination index, b_g - an item difficulty index, and c_g - a pseudo guessing parameter. Using the notation $P(u_g=1|\theta_i)$ to represent the probability of a correct response to item g given an examinee of ability level θ_i, Birnbaum's three parameter model states that:

$$P(u_g=1|\theta_i) = c_g + (1 - c_g)\ [1+\exp\ (-1.7a_g\ (\theta_i - b_g)\ )]^{-1}$$

This relationship between θ_i and $P(u_g=1|\theta_i)$ is illustrated in Figure 2.

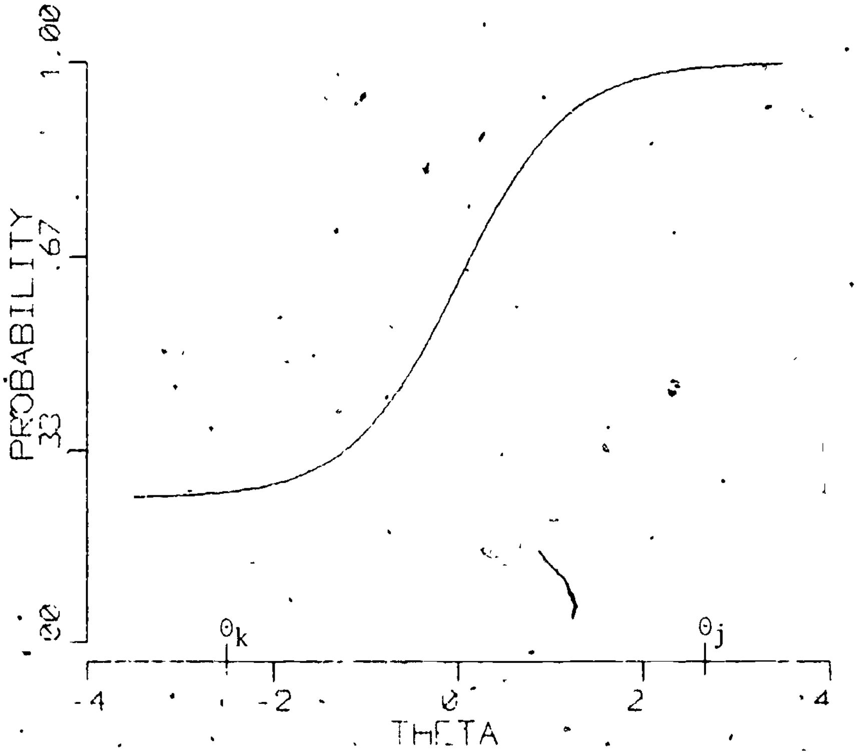

Figure 2: A hypothetical item characteristic curve

The probability of a correct response given a specific ability level increases monotonically as true ability increases. For example, an examinee with a high true ability, e.g. θ_j, has a high probability of responding correctly $[P(u_g=1|\theta_j)\rightarrow 1.0]$. Conversely, an examinee of low true ability, e.g. θ_k, has a low probability of responding correctly; approaching the lower asymptote of the curve, c_g.

The inflection point of the curve, b_g, is referred to as the item difficulty parameter in that it indicates the relative position of the curve along the θ axis, The more the curve is positioned to the right, the more ability is necessary for an examinee to have a good probability of a correct response. The slope of the curve at b_g helps define a third parameter, a_g. This value, referred to as the discrimination parameter, indicates the power of the item to separate examinees of close but unequal levels of ability. Although the item parameters and θ are on a common metric, these item parameters described characteristics of the item independently of the examinee group. Full explanations and development of this and other mental measurement models can be found in Jensema (1972) and in Lord and Novick (1974).

Latent trait theory has been used to identify biased items (Green and Draper, 1972; Lord, in press; Rudner, 1977a). In an early study, Green and Draper had used observed total scores as estimates of examinees' abilities, θ_i's, and the proportions of examinees responding correctly at each total score level as estimates of $P(u_g=1|\theta_i)$. Their procedure called for plotting estimates icc's for each item separately for each culture group and comparing the plots,

By this and other latent trait theory approaches, an item is unbiased if examinees of the same ability level, but of different cultural affiliations, have equal probabilities of responding correctly. That is, an item is unbiased if the estimated icc's obtained from the various culture groups are identical. As an example of a biased item, consider the two hypothetical curves shown in Figure 3. These curves are based on responses by two different culture groups to the same item. Total observed scores are used as estimates at Θ_i and proportions of examinees responding correctly are used as estimates of $P(u_g=1|\Theta_i)$. The curves are not identical, since the location parameters for the two curves are not equal. Such an item can be considered biased in that often examinees of the same ability level, e.g. X = 58%, but from different culture groups, do not have similar proportions of correct responses.

While this approach is appealing, total observed scores are directly incorporated and quantification of the degree of item bias is difficult (an eyeballing procedure is used to identify a "very biased item").

Rather than using total observed scores as estimates of Θ_i and proportions as estimates for $P(u_g= 1|\Theta_i)$, more accurate values can be obtained using one of the recent methods of parameterization (Urry, 1975; Wingersky and Lord, 1973). During parameterization, the metric used for the Θ scale is defined by the ability variance in the examined sample. In order to compare parameters obtained from two different examinee groups, the obtained values must be equated. Lord and Novick (1974, Chapter 16.11) and Rudner (1977b) have shown that this can be accomplished by computing the regressions of the parameter values based on one group of examinees on the parameter values based on the other group of examinees. The equated icc's will be identical when the restrictions of the model are met. That is, when the measure:

(1) is unidimensional

(2) contains locally independent items.

(3) has error-free parameter estimates.

Rudner (1977a) has refined the procedure used by Green and Draper to

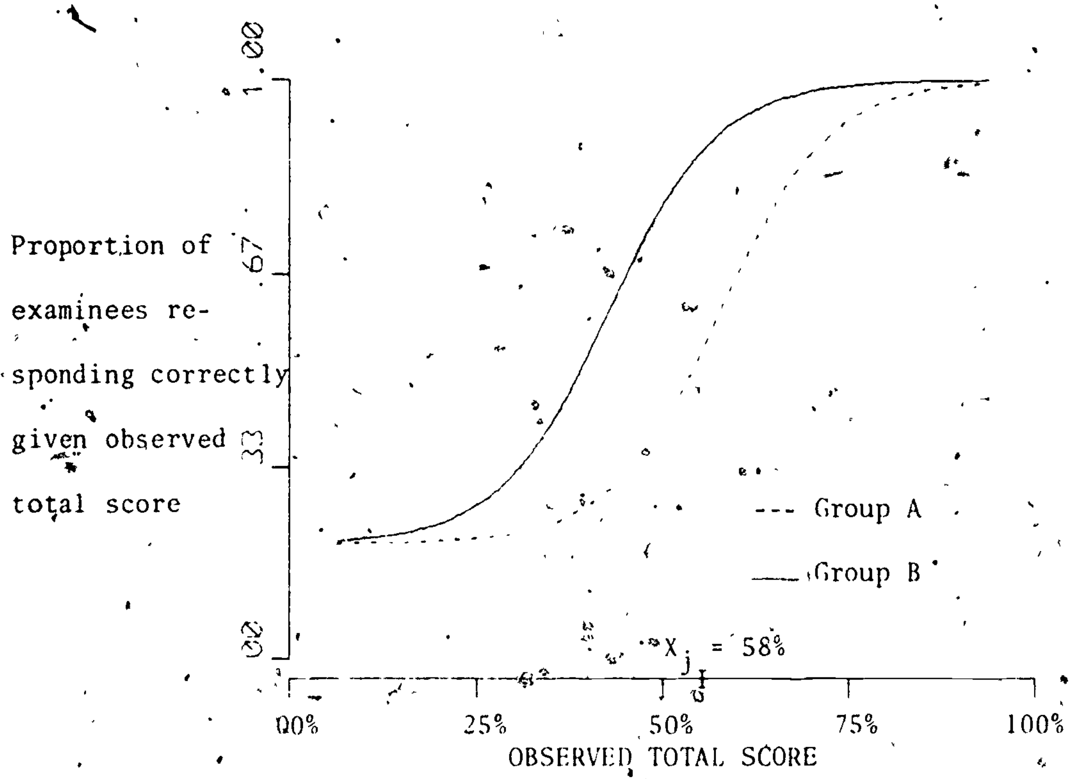

Figure 3: Two hypothetical response distributions

identify biased items by incorporating equated icc parameter values. The area between pairs of equated icc's is used to indicate the relative amount of aberrance for each item and eyeballing of the equated icc's is employed to provide additional information as to the nature of the aberrance. Lord (in press) has employed an asymptotic significance test based on the summed variance-covariance matrices of the equated a_g and b_g parameter estimates to test for significant differences between pairs of equated icc's.

Factor Analytic Approaches

In factor analysis, underlying factors (i.e., dimensions or traits) are hypothesized and the correlations of each variable with the hypothesized factors are computed. In an achievement test, each item is treated as a variable. Such an analysis could be conducted twice using examinees from two different cultural backgrounds. Ideally, the two separate groups of examinees would yield similar sets of item-trait correlations (factor loadings). Different sets of factor loadings would indicate that the two groups are not responding to the items in the same manner. Such a test would be considered biased in that it appears to measure a different trait across groups. The items exhibiting the most bias would then be those with the largest differences in factor loading.

The general model for this type of factor analysis is

$$\underline{y} = A\underline{f} + \underline{e}$$

where $\underline{y}$ is a vector of subject responses

A is a matrix of factor loadings

$\underline{f}$ is a vector of factor variables (locations)

$\underline{e}$ is a vector of residual or error terms

From $\underline{y}$, values of A, $\underline{f}$, and $\underline{e}$ are determined.

Green and Draper (1972) and Green (1976) suggest an inner-group factor analysis model based on the inner-battery factor analysis approach offered by

Tucker (1958). In this inner-group model, the item variance is partitioned into: (1) factors common to each subgroup; (2) factors specific to subgroups; and (3) residual or error variance. With this model one can determine the proportion of item variance accounted for by a given subgroup. An item, then, is unbiased when this proportion is small and biased if a large proportion of variance is attributable to culture-specific sources.

Merz (1973, 1976a) developed an alternate approach which incorporates factor scores and analysis of variance. In this approach, the item responses for the groups are combined, factor analyzed, and factor scores for each examinee on each factor computed. These factor scores are then subjected to an analysis of variance, with group membership being the independent variable. Where significant mean differences are found in factor scores, the factor is classified as biased. Biased items are defined as those with high factor loadings on a biased factor.

These approaches are appealing in that they deal with the underlying latent traits (true abilities) of the examinees. Green and Draper's approach is particularly appealing in that variance is partitioned into culture-specific and culture-common sources. Merz's approach has an advantage in that variance caused by factors such as socio-economic status, IQ, and sex can be partialled out. However, these procedures are not without some conceptual as well as practical limitations.

The first step in factor analysis is the computation of the inter-variable correlations matrix. To obtain stable correlations--to avoid capitalization on change--one needs a large number of subjects; the general rule of thumb is at least ten subjects per variable, a figure often ignored in practice.

Assuming a sufficient number of subjects, there is a question as to which type of correlations to use. In analyzing items, one usually deals with dichotomously scored variables and either the phi (product-moment) or the tetrachoric correlation is employed. However, the phi correlation is limited in that it is highly sensitive to item difficulties, and the tetrachoric correlation, though it estimates what the value of the inter-item correlation would be if the items were continuous variables, is notoriously unstable. In fact, Nunnally (1967, p. 124) emphatically states that tetrachoric correlations cannot be used in factor analysis.

Regardless of which type of correlation is used, there are additional problems. As Nunnally (1967) points out, ". . . for a group of variables to clearly define a number of factors, there must be a wide range of correlations" (p. 256). In correlating items, especially dichotomously-scored items, the average correlation is typically low. ~~Thus,~~ it usually is not possible to obtain a clear factor structure when factor analyzing test items.

Finally, in factor analysis many decisions need to be made by the researcher. Which procedure? How many factors to extract? Which rotational scheme to use? Different decisions can lead to different results. Thus while the factor analytic approaches are appealing, in practice they may be difficult to apply.

Distractor Response Analysis

Some of the chi-square, item difficulty regression, item characteristic curve-theory, analysis of variance, and factor analysis approaches incorporate total test scores either directly or indirectly. This can pose a problem when the total scores do not represent accurately the abilities of the examinees, as would be expected in a very biased test.

Veale and Foreman (1975, 1976) recommend investigating the distractor response distribution for various cultural groups in an approach not dependent upon this assumption. Should one group be overly attracted to a particular distractor in comparison to a second group, there may be a biasing characteristic of the item attracting them away from the correct response. Bias is thus defined as characteristics of an item which cause a distortion in the item p-value for a cultural group.

Consider the choice distribution illustrated in Table 1. Observed frequencies appear in the cells and expected frequencies appear in the upper right hand corner of each cell. A disproportionate number of members of Group 2 were attracted to Distractor 1 (the response frequencies can be shown to be disproportionate by the use of a chi-square test). It may be argued that some characteristic of Distractor 1 caused a substantial number of members of Group 2 to select this distractor over the correct alternative. Hence some characteristics of the item may have caused a distortion in the group p-value.

Table 1

A Hypothetical Item Distractor Choice Distribution
Frequency of Selection

		Distractor 1	Distractor 2	
Group	1	60 40	40 60	100
	2	60 80	40 20	100
		120	80	200

To obtain a global picture of an item's behavior, Veale and Foreman compute several statistics on each item. These include:

(1) a chi-square to test the hypothesis that the conditional probabilities of individuals missing the item by selecting a particular distractor given their cultural group (foil pull indices) are equal across cultural groups;

(2) Cramer's V as a measure of "cultural variation" to determine the extent of departure from the hypothesis tested above;

(3) Goodman-Kruskal measures of index groups by distractor association;

(4) supplementary item statistics for each cultural group including z-tests for testing deviations from random guessing, p-values, point biserial correlations, and chi-square tests for gauging deviations from uniform distractor response distribution.

These supplementary statistics help discriminate between desirable and undesirable items. For example, an item may show low cultural variation among the distractors and have highly different point biserial correlations between cultural groups. Such an item would appear to work well with one group and poorly with another. This information, coupled with the variance in the distractor distributions, would probably lead either to elimination or revision of the item.

While directly sensitive to bias in item distractors, this approach is only indirectly sensitive to other sources of bias such as those in the item stem, directions, or subject matter. If one suspects that item bias is most often caused by bias in the distractors, this limitation is not a serious one. Further, by supplementing this approach as Veale and Foreman suggest, it is possible to obtain a holistic view of the behavior of the aggregate item and its-constituent distractors.

Like the earlier chi-square and analysis of variance approaches, distractor response analysis requires a large number of inferential tests and the consequent probability of committing Type I errors must be realized.

Discussion and Summary

Several approaches toward the identification of biased items have been presented with their rationale and apparent advantages and limitations. Comments have also been made regarding the use of a large number of inferential tests, the assumption of an unbiased total score, and the use of outlier analysis (see Table 2). In practice, depending on the purpose of the study and the initial item pool, these limitations may be inconsequential.

The practitioner must first delineate the purpose to which such approaches are to be applied. One purpose is to debias an instrument during its development. The degree of item bias (indicated by the magnitude of a residual, area, factor loading, X^2 or F) can be considered along with professional judgments of item difficulty indices, item discrimination indices, and factor loadings to determine which items are to be retained and dropped. In such instances, it is usually better to drop an item falsely suspected of being biased than to retain a truly biased one. Here, the limitions caused by inflated alpha errors may be moot in the chi-square, distractor response analysis, and analysis of variance approaches.

On the other hand, these techniques can be used to identify trends in biased items. That is, biased items can be pooled and attempts made to identify salient characteristics (see Rudner, 1977b). In such instances, one would want a more conservative identification procedure. The transformed item difficulties and the item characteristic curve theory approaches are well-suited for this in that the confidence band can be narrowed or widened as desired.

Table 2

Some Salient Characteristics of the Different Approaches

(Part I)

	Analysis of Variance	Transformed Item Difficulties	Chi-Square	Item Characteristic Curve Theory
Major literature	Cardall and Coffman (1964)	Angoff (1972)	Scheuneman (1975, 1976)	Rudner (1977a)
Operational definition	Significant analysis of variance item by group interaction	Differential relative item difficulty	Proportion of correct responses to an item is unequal for members of different groups within the same total score category	Probability of a correct response for a given true ability is unequal for examinees from different groups
Dependence on total score being valid	Indirectly	Indirectly	Directly	No
Computational ease	Difficult	Easy	Easy	Difficult
Ease of conceptual understanding by lay-people	Medium	Easy	Easy	Difficult
Applicability to criterion referenced tests	Low	Low	Low	Low

Table 2

Part I (Continued)

	Analysis of Variance	Transformed Item Difficulties	Chi-Square	Item Characteristic Curve Theory
Applicability to easy (difficult) items	Medium (medium)	Medium (medium)	High (low)	Medium (medium)
Applicability to more than two independent and/or interdependent cultural groups	High	Low	Medium[1]	Low
Applicability to multiple choice items (non-multiple choice items)	High (high)	High (high)	High (high)	High (high)

[1] By appropriately defining specific group membership; e.g., black females, as the independent variable

Table 2

Some Salient Characteristics of the Different Approaches

(Part II)

	Factor Analysis	Factor Score	Distractor Response Analysis
Major literature	Green (1976); Green & Draper (1972)	Merz (1973, 1976)	Veale & Foreman (1975, 1976)
Operational definition	Large proportion of item variance is group specific	High loading on a factor which yields unequal group mean factor scores	Characteristic(s) of the item distorts group item p-values
Dependence on total score being valid	Indirectly	Directly[1]	No
Ease of conceptual understanding by lay-people	Difficult	Difficult	Easy
Applicability to criterion referenced tests	Low	Low	High
Applicability to easy (difficult) items	Medium (medium)	Medium (medium)	Low (high)
Applicability to more than one independent and/or interdependent cultural groups	Medium[2]	High	Medium[2]
Applicability to multiple choice items (non-multiple choice items)	High (high)	High (high)	High (no)

[1] In computing factor scores.

[2] By appropriately defining specific group membership; e.g., black females, as the independent variable

Often the intended audience will be a deciding factor in determining which approach to use. Here the distractor response analysis, chi-square, and transformed item difficulties approaches have a distinct advantage since they are computationally and conceptually easy and can be readily explained to the layperson.

One may wish to develop a measure that is simultaneously unbiased for three cultural groups such as white, black, and Chinese Americans. An extension of this involves interdependent culture groups such as male-female and white-black comparisons. Such interactions and simultaneous comparisons can be analyzed directly by either the analysis of variance or factor score approaches. The chi-square, distractor response analysis, and factor analysis approaches can be adapted readily for such an analysis by defining group membership appropriately. The transformed item difficulties and icc theory approaches can also be applied but only by using several pairwise comparisons.

One final consideration is applicability to criterion referenced tests. Ideally, the items of such measures are designed to be sensitive to growth, rather than to differences among students. Examinees who have not mastered an objective are expected to respond erroneously while those who have met the criterion level are expected to respond correctly. Thus one cannot expect the large variance of total scores (occasionally coupled with a normality assumption) required of all the approaches other than distractor response analysis. Therefore, if one is interested in analyzing items in a true criterion referenced test, distractor response analysis appears to be the only alternative presently available.

In summation, there is no one approach which appears best suited for all situations. Of the approaches, the distractor response analysis and chi-square approaches are the most communicable--a distinct advantage in explaining a debiasing investigation to the lay person. In actually pinpointing the source of bias, distractor response analysis is particularly useful because it alone identifies which response alternative is the cause of aberrance. In addition, distractor response analysis is uniquely applicable to true criterion referenced tests. In terms of statistical adequacy, the icc theory approach is appealing in that it is a true score model making no assumption about the accurancy of p-values or individual total scores.

REFERENCES

Anastasi, A. Psychology, psychologists and psychological testing. American Psychologist, 1967, 22, 297-306.

Anastasi, A. Psychological Testing (4th Ed.). New York: MacMillan, 1976.

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

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

Angoff, W. H., & Modu, C. C. Equating the scales of the Prueba de Aptitud Academica and the Scholastic Aptitude Test. New York: College Entrance Examination Board, 1973.

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, Statistical Theories of Mental Test Scores. Reading, MA: Addison-Wesley, 1968, Chapts. 17-20.

Brogden, H. E. Variation of test validity with variation in the distribution of item difficulties, number of items, and degree of the intercorrelation. Psychometrika, 1946, 11, 197-214.

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

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

Cleary, T. A., & Hilton, T. L. An investigation into item bias. Educational and Psychological Measurement, 1968, 8, 61-75.

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

Comprehensive Test of Basic Skills, Form S. Technical Bull. 1, Monterey: CTB, McGraw-Hill, 1974.

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

Darlington, R. B. Is culture fairness objective or subjective? Paper presented at the annual meeting of the American Educational Research Association, New Orleans, 1973.

Darlington, R. B. A defense of "rational" personnel selection and two new methods. Journal of Educational Measurement, 1976, 13, 31-41.

Duncan, D. B. Multiple range and multiple F-tests. Biometrika, 1955, 11, 1-4.

Duncan, D. B. Multiple range tests for correlated and heteroscedastic means. Biometrika, 1957, 13, 164-176.

Eagle, N., & Harris, A. S. Interaction of race and test on reading performance scores. Journal of Eduational Measurement, 1969, 6, 131-135.

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

Eells, K., Davis, A., Havighurst, R. J., Herrick, V. E., & Tyler, R. W. Intelligence and Cultural Differences. Chicago: Unviersity of Chicago Press, 1951.

Einhorn, H. J., & Bass, A. R. Methodological considerations relevant to discrimination in employment testing. Psychological Bulletin, 1971, 75(4), 261-269.

Fitzgibbons, T. J. The use of standardized instruments with urban and minority-group pupils. New York: Harcourt Brace Jovanovich Test Department, 1971.

Green, D. R. Biased tests. Monterey: CTB/McGraw-Hill, 1971.

Green, D. R. Racial and ethnic bias in test construction. Monterey: CTB/McGraw-Hill, 1972.

Green, D. R. Reducing bias in achievement tests. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1976.

Green, D. R., & Draper, J. F. Exploratory studies of bias in achievement tests. Monterey: CTB/McGraw-Hill, 1972.

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

Guilford, J. P. Psychometric methods. New York: McGraw-Hill, 1954.

Hicks, M. M., Donlon, T. F. & Wallmark, M. M. Sex differences in item responses on the Graduate Record Examination. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, April 1976.

Hoepfner, R., & Strickland, G. P. Investigating test bias. Los Angeles: Center for the Study of Evaluation, University of California, 1972.

Jensema, C. J. An Application of latent trait mental test theory to the Washington pre-college testing battery. Unpublished Doctoral Dissertation, University of Washington, 1972.

Jensen, A. P. An examination of culture bias in the Wonderlic Personnel Test. Arlington, VA: Eric Clearinghouse, 1973 (ERIC Document Reproduction Service ED 086 726).

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

Linn, R. L., & Werts, C. E. Considerations for studies of test bias. Journal of Educational Measurement, 1971, 8(1), 1-4.

Lord, F. M. A theory of test scores. Psychometric Monograph Number 7. Princeton: Educational Testing Service, 1952.

Lord, F. M., & Novick, M. R. Statistical Theories of Mental Test Scores (2nd Ed.). Reading, MA: Addison-Wesley, 1974.

Lord, F. M. A study of item bias using item characteristic curve theory. Proceedings of the Third Congress of Cross-Cultural Psychology, Tilburg, Holland, in press.

Merz, W. R. Factor analysis as a technique in analyzing test bias. Paper presented at the annual meeting of the California Educational Research Association, Los Angeles, 1973.

Merz, W. R. Estimating bias in test items utilizing principle component analysis and the general linear solution. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1976a.

Merz, W. R. Test fairness and test bias: a review of procedures. Paper presented at the Office of Education Invitational Conference on Achievement Testing of Disadvantaged and Minority Students for Educational Program Evaluation, Reston, VA, May 1976b.

Munday, L. A. Measurement for equal opportunity. The Counseling Psychologist, 1970, 2(2), 93-97.

Ozenne, D. G., Van Gelder, N. C., & Cohen, A. J. Emergency school aid act (ESAA) national evaluation achievement test restandardization. Santa Monica, CA: System Development Corporation, 1974.

Rasch, G. Probabilistic Models for Some Intelligence and Attainment Tests. Copenhagen: Denmarks Paedoggiogishe Institute, 1960.

Ruch, G. M. The Objective or New-Type Examination. Chicago: Scott Foresman and Company, 1929.

Rudner, L. M. 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, April 1977a.

Rudner, L. M. Item Bias with Deaf and Hearing Examinees. Paper presented at the annual convention of American Instructors of the Deaf, Los Angeles, June 30, 1977b.

Rudner, L. M. A closer look at latent trait parameter invariance. Paper presented at the annual meeting of the New England Educational Research Organization, Manchester, N.H., May 5, 1977c.

Schmidt, F. L., & Gugel, J. F. The Urry Item Parameter Estimation Technique: How Effective? Paper presented at the American Psychological Association Convention, Chicago, August 1975.

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

Scheuneman, J. A procedure for evaluating item bias in the absence of an outside criterion. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1976.

Siegel, S. Nonparametric Statistics for the Behavorial Sciences. New York: McGraw-Hill, 1956.

Strassberg-Rosenberg, B., & Donlon, T. F. Context influences on sex differences in performance and aptitude tests. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, D.C., 1975.

Thorndike, H. L. Concepts of cultural fairness. Journal of Educational Measurement, 1971, 8, 63-70.

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

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

Urry, V. W. A Monte Carlo investigation of logistic mental test models. Unpublished Doctoral Dissertation, Purdue University, 1970.

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

Veale, J. R., & Foreman, D. I. Cultural validity of items and tests: A new approach. Score Technical Report, Iowa City, Iowa: Westinghouse Learning Corporation/Measurement Research Center, 1975.

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

Williams, R. L., Mosby, D., & Hinson, V. Critical issues in achievement testing of children from diverse ethnic backgrounds. Paper presented at the Office of Education Invitational Conference on Achievement Testing of Disadvantaged and Minority Students for Educational Program Evaluation, Reston, VA, May 1976.

Wingersky, M. S., & Lord, F. M. A computer program for estimating examinee ability and item characteristic curve parameters when there are omitted responses. (RM 73-2.) Princeton: Educational Testing Service, 1973.