DOCUMENT RESUME

ED 268 160 TM 860 185

AUTHOR	Bleistein, Carole A.
TITLE	Application of Item Response Theory to the Study of Differential Item Characteristics: A Review of the Literature.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RM-86-3
PUB DATE	Jan 86
NOTE	34p.
PUB TYPE	Information Analyses (070)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Black Students; Comparative Analysis; *Culture Fair Tests; Literature Reviews; *Racial Differences; Research Problems; Sampling; Secondary Education; Statistical Analysis; *Test Bias; Testing Problems; Test Validity; White Students
IDENTIFIERS	Canada; *Three Parameter Model

ABSTRACT

Research on assessing the cultural fairness of individual test items is reviewed, with emphasis on Birnbaum's three-parameter logistic model. As defined in this review, differential item characteristics are exhibited when examinees from one group have a lower probability of answering correctly than do examinees of equal ability from another group. An item is assumed to be free of bias if the probability of responding correctly, given total score, is the same for all subpopulations. Other major methods of identifying differentially performing items and their limitations are briefly summarized: analysis of variance; transformed item difficulties; chi-square; and factor analysis. Twelve studies are reviewed. Subjects generally included black and white students; in some cases, hearing impaired students and Canadians were included. Comments by Petersen, Lord, and Ironson are also summarized. It is concluded that the three-parameter model's advantages include the relationship of the probability of a correct response to underlying examinee ability level and item characteristics; the model's ability to take guessing into account; and the fact that item parameters are sample-independent and invariant across groups. The model's limitations are also described; it is often difficult to use and impossible to apply. Sample sizes must be at least 1,000 and there should be at least 40 items to be analyzed. (GDC)

RM-86-3

RESEARCH MEMORANDUM

APPLICATION OF ITEM RESPONSE THEORY TO THE STUDY OF DIFFERENTIAL ITEM CHARACTERISTICS

A REVIEW OF THE LITERATURE

Carole A. Bleistein

Educational Testing Service
Princeton, New Jersey
January 1986

It is widely recognized that test performance may be influenced by variables other than ability in the trait being measured and, to the extent that such influence occurs, the validity of the test may be compromised. Researchers and test developers concerned with culture-fair testing have addressed this issue through a variety of models; some designed to study the fairness of the test as a whole, others to assess the fairness of individual items within the test. It is research devoted to the latter that is the concern of the present paper.

Throughout this text, the term "differential item characteristics" is used to denote performance differences that may be attributable to subgroup membership. The following conceptual definition of differential item characteristics is offered:

> An item is exhibiting differential item characteristics if examinees from one group have a lower probability of answering the item correctly than do examinees of equal ability from another group.

For purposes of the present paper, it is assumed that an item is free of differential characteristics if the probability of responding correctly, given total score, is the same for all subpopulations studied. It is recognized that this reliance on an internal criterion is not without problems in the sense that the criterion itself may not be free of differential characteristics. However, total test score is probably the best measure of ability on the trait being tested that is available and it is the measure used in the papers to be reviewed.

The major methods for identifying differentially performing items include analysis of variance (Cleary and Hilton, 1968), transformed item

difficulties (Angoff, 1972), chi-square approaches (Scheuneman, 1979), factor-analytic approaches (Green and Draper, 1972), and item response theory approaches (Lord, 1952, Rasch, 1960 and Birnbaum, 1968). The purpose of this paper is to describe applications of the item response theory (IRT) approach to the analysis of differential performance characteristics and, therefore, a full description of the other methodologies is beyond the scope of the present paper. The interested reader is referred to the sources cited. However, since item response theory addresses some of the limitations of the other methodologies in terms of identifying differentially performing items, such limitations are briefly summarized below. The limitations of item response theory itself will be discussed later in the paper.

- o Analysis of variance. Significant item-by-group interactions have been accepted as evidence of differential item characteristics. However, Hunter (1975) has shown that such interaction will always occur when two groups differ in ability, and the approach is sample-dependent.
- o Transformed item difficulties. Pairs of P-values converted to normal deviates, are plotted and items deviating greatly from the line of best fit are considered to be exhibiting differential performance characteristics. This approach also defines differential performance as an item-by-group interaction and is, therefore, subject to the same limitation as analysis of variance, that is, it is sample-dependent.

- o Chi-square approaches. An advantage of chi-square approaches is that the sample is blocked on the basis of ability and the presence of differential item characteristics is inferred if proportion of correct responses, given ability, is not the same for the two groups. However, expected frequencies are affected by total score distributions, and to the extent that such distributions are different for the two groups, chi-square will be inflated (Rudner and Convery, 1978).
- o Factor analysis. An advantage is that ability level of examinees is dealt with, however, as Rudner, Getson and Knight (1980) point out, results may depend, to a significant degree on user decisions in terms of type of correlation matrix, type of factor analysis, number of factors and type of rotation to be used.

The literature to be reviewed is confined primarily to studies which have applied the three-parameter logistic model to the assessment of differential item characteristics, but since much of the early work in the field makes use of the Rasch one-parameter model, one such report has been chosen as representative.

Wright, Mead, and Draba (1976) state that the Rasch model, begins with assumptions similar to those of previous differential item characteristics work and its procedures are extensions of traditional item analysis which permit the identification of differentially performing items for individuals as well as groups. They suggest using only internal criteria because of the difficulty of constructing external criteria that are free of differential performance, and state that only by a logistic transformation of item

difficulties is it possible to compare item difficulties across groups without contamination by within-group ability. The asymptotic estimates of the variance of parameter estimates resulting from maximum likelihood techniques enable us to: (1) identify differentially performing items by placing them on a common metric, (2) to determine whether differential performance is responsible for poorly fitting items and (3) to determine expected residual variance in cases where the data do appear to fit the model and, thus, aid in diagnosing the sources of invalidity.

In order to meet the conditional requirements of a good measurement model (i.e., persons with higher θ (ability) have higher P(success), P(success) is higher for easy items for all people, and P(success) is a consequence of θ and difficulty level), Wright, et al. consider only tests comprised of homogeneous items. If the above conditions are met, they believe that the Rasch model specifying ability and item difficulty should fit and, consequently, we can estimate each of the parameters independently of the other. They claim that raw score is a sufficient statistic for ability and that number of persons responding correctly is a sufficient statistic for item difficulty and, therefore, that the model is consistent with traditional item analysis with the exception that the indices of traditional item analysis are sample-dependent. All that is necessary to overcome the consequent distortion of the proportion metric is to adjust the ability distribution to obtain sample-independent estimates of item difficulty.

The Rasch model is justified by Wright, et al. on the grounds that if there are other parameters, consistent estimators for them do not exist and

attempted applications of multiparameter models have not been successful. The logistic form provides statistically independent item- and person-parameters and allows us to make statements about person-item interactions, thereby, facilitating analysis of differential item characteristics. Wright, et al. suggest the construction of a variety of indices based on linear analysis of residuals to diagnose problems such as speededness and guessing.

Pine (1977) points out that by definition, for an item to be free of differential item characteristics, it must have the same item response function for all subgroups, and adds that this requires that three parameters, item difficulty, discrimination and pseudo-chance, be invariant up to a linear transformation across subgroups. He specifies a linear transformation to resolve problems of different ability distributions of various subgroups. Pine, too, restricts his discussion to homogeneous tests and cites Lord and Novick (1968) as having shown that parameter invariance holds as long as the test is unidimensional across subgroups. He states that because of this property, factor analyzing the inter-item correlation matrix for each subgroup and showing that a single dimension accounts for nonrandom variance should be adequate to assess differential item characteristics in a test. He suggests that the linear relationship of item parameters for unidimensional tests be assessed by separately plotting each of the parameters for one group against those of the other group and testing for departures from linearity. Such a comparison of item parameters will enable us to detect which items cause the nonlinear item-by-group interactions and the perpendicular distance from the line of best fit to the

plotted points may be used as the measure of differential item characteristics.

Pine points out the similarity between his method and Angoff's delta plot method, the essential difference being that the delta plot method uses classical parameters which are not linearly related across subgroups and which, therefore, may lead to an artifactual detection of differential item characteristics. In addition, he criticizes the difficulty index of classical test theory on the grounds of confounding difficulty with discrimination and guessing in contrast to item response parameters.

His data consisted of responses to 75 objective items by 58 blacks and 168 whites from a Minneapolis high school. Pine acknowledges that his sample size is smaller than is generally considered adequate, but believes it serves well enough for demonstration purposes. Because his sample contained only 58 Blacks, he used a 45-item subtest to do a principal-axis factor analysis on the tetrachoric correlation matrices for each group to test the unidimensionality assumption, with the result that the first eigenvalue for both groups was very large, implying unidimensionality. He also found a .97 coefficient of congruence between the factor loadings for Factor 1 in the two subgroups. He plotted the difficulties for the two groups to assess the degree of linearity and found a correlation of .86 which he took as evidence that the test was free from differential item characteristics. He does not explain how he obtained the difficulty parameter, but does state that discrimination and guessing parameters were not estimated because of their unreliability with samples this small. He also noted that for those items which substantially deviate from linearity, an index of differential item characteristics should be determined.

Among the criteria for an improved approach to the study of differential item characteristics, developed by Rudner (1977), are: (1) sensitivity only to group differences in the trait being tested, i.e., the approach should be relatively insensitive to factors other than differential item characteristics which might be affecting performance, such as ability differences between the groups, (2) total observed score should not be assumed to be a valid ability measure, (3) the approach should allow quantification of the degree of differential item characteristics and (4) it should apply to items which vary widely in difficulty.

Rudner suggests that latent trait theory meets the above criteria. His study was conducted using two approaches. In the first, his sample consisted of pseudo-culture groups obtained by randomly dividing approximately 2,600 hearing-impaired high school students into two groups of differing mean ability and treating them as if they were different cultural groups. In the second approach, he used the total 2,600 hearing-impaired students as one cultural group and approximately 1,600 West Coast public high school students as the other. All groups were administered the 1973 Stanford Achievement Test-Reading Comprehension Subtest. The application of item response theory consisted of estimating parameters on each group separately, equating the scales and calculating the area between the item response functions to indicate the degree of differential item characteristics.

The pseudo-cultural group comparison identified two items exhibiting differential performance, probably due to poor parameter estimation. Their b (difficulty) values were extremely high for one group where few examinees

had high enough θ's to answer the items correctly. Consequently, parameter estimates were based on a small number of subjects for these items. Comparison of the two diverse culture groups identified five items favoring hearing examinees and one item favoring the non-hearing. He also found that directionality of differential performance could not always be determined since one item favored both low ability hearing-impaired and high ability hearing examinees, i.e., the item characteristic curves crossed. Two items could not be parameterized due to very low point biserials, indicating poor relationship between ability and the probability of answering the item correctly.

Rudner concluded that the IRT approach is the one which best meets his criteria, but stresses the importance of eyeballing the item response functions in addition to computing distances between them in assessing differential performance.

Lord (1977), in his attempt to determine whether the Scholastic Aptitude Test measures the same thing for blacks and whites and whether some items should be removed so that it measures appropriately in both groups, applied IRT to the study of differential item characteristics. His sample consisted of 2,250 each of white and black high school students with scores on the April 1975 SAT (85-item Verbal section). He stressed that the P value is not a measure of item difficulty since it is group-dependent and that parameter invariance is the outstanding advantage of item response theory. Any difference between item response functions is indicative of some kind of differential item characteristics. Lord first used the LOGIST program to estimate θ's and item parameters for each group separately. These

parameters, however, cannot be directly compared since the origin and unit for measuring ability can't be determined from the data. For this reason, the b parameters of one group were plotted against those of the other group. The straight line fitted to the points was used to put all item parameters on the same scale. Lord used an asymptotic significance test to test the null hypothesis that item response functions for both groups were the same.

He found that the item response functions for 46 of the 85 items were significantly different at the .05 level. After eliminating items with significant differences beyond the .15 level, (leaving 32 items), he combined groups and reran LOGIST for the reduced 32-item subtest so that ability parameters for both groups were on the same scale. The entire first step was repeated, treating the estimated θ's as given. He again performed the asymptotic significance tests, rejecting the null hypothesis for 38 of the 85 items.

Lord also combined his black and white groups into a total group and then randomly divided the total group into what he termed "blue" and "red" groups on which he repeated the procedure detailed above. He pointed out that the 85 items should be rectangularly distributed over the range of significance levels since the groups were formed randomly. What he found was very close to this, that is, his asymptotic significance tests are very close approximations. He found that one-third of the items really did have different item response functions for the black and white groups. Despite this, he concluded that the test does measure approximately the same skill for blacks and whites, although some of the items act differently for the two groups.

Haebara (1979), criticizes previous differential item characteristics work in which the possible differences in the guessing parameters for different subpopulations are not taken into account. His proposed approach utilizes information contained in all three parameters and includes statistical tests of parameter invariance. His sample consisted of 400 each black and white students, with males and females equally represented, drawn from schools located in various parts of the country, and his instrument was the 60 5-choice items of the Reading Comprehension subtest of the Tests of Achievement and Proficiency.

In the absence of an external criterion, Haebara emphasizes the importance of the assumption that the test as a whole is free from differential item characteristics. His method of detecting differential characteristics was to examine parameter invariance across subgroups on the basis of approximate standard errors of the maximum likelihood estimates of the IRT parameters. Haebara stressed the importance of the guessing (c) parameter in assessing differential performance. Rudner's method of equating a (discrimination) and b (difficulty) parameters ignored the information contained in c (pseudo-chance). In contrast, Haebara's procedure was to estimate θ's and item parameters separately for each subgroup (using LOGIST), to compute approximate asymptotic standard errors of the estimate of c, and to test the significance of the difference between the c's. For items not found to be performing differentially with respect to c, a common c value was assigned for both groups. Haebara lists three ways in which the common c value may be determined: (1) using the total groups and items not already rejected, use LOGIST to reestimate c, (2) weight the

average of the original estimates by their respective relative precision or (3) use estimates from the lower performance group because it has more examinees in the lower tail of the ability distribution. This is in contrast to Rudner's approach of replacing values of c in the lower ability group with those from the more able group.

Haebara chose the third approach to save computer time for reestimating a and b parameters for the black group, which would otherwise have been necessary. The a and b parameters were then equated and the first three steps were repeated from them. Haebara, too, uses approximate asymptotic significance tests, and specifies that they are based on approximations of the sampling distributions of maximum likelihood estimates and so are not exact tests. The approximate significance tests were performed first on the invariance of the c's. Approximate standard errors of the c's were computed for each subgroup. Items rejected through this procedure are performing differentially with respect to c. These items were not tested on invariance of a's and b's, but their item response functions were inspected. Haebara then proceeded to equate the a's and b's and to run the significance tests on the equated a and b parameters.

He found that the white group performed better and also had higher c values. In addition, there were 15 items identified as performing differentially with respect to a and 17 with respect to b.

The necessity for equating a and b parameters was explained by Dorans (1979) in his discussion of the need for a common metric in differential item performance studies. It was his contention that if this need was overlooked, any of the conclusions could be invalid. He stressed that,

while the choice of metric may be arbitrary, it is essential that the chosen metric be maintained across subgroups. This is seen as particularly important in studies that depend on assessing parameter invariance as a means of detecting differentially performing items, since that property only holds if a common metric is used for all individuals. If parameters for subgroups are on different metrics, apparent differential performance will be artifactual.

Ironson and Subkoviak (1979) address the issue of evaluating the validity of differential item characteristic methods by assessing the extent to which four procedures (transformed item difficulty, item discrimination differences, chi-square and IRT) agree or disagree in the identification of differentially performing items. To this end they analyzed data from six subtests of the 1972 National Longitudinal Study (vocabulary, picture-number, reading, letter groups, mathematics and mosaic comparisons) administered to 17,726 12th-grade students, from which they selected a sample of 1,691 blacks and 1,794 whites. Their findings indicated that the highest average differential item characteristic indices occur in the vocabulary and reading subtests with transformed item difficulty, chi-square and IRT approaches. The discrimination difference index seemed to be unrelated to the subtest analyzed. The largest correlation between methods (.485) was between chi-square and IRT and the discrimination approach didn't correlate significantly with any of the other methods. The correlations between the IRT method and a traditional vs. non-traditional classification of the subtests (i.e., vocabulary, reading and math vs. picture-number, letter groups and mosaic comparisons) were the highest, followed by

chi-square and transformed item difficulty. Ironson, et al. concluded that there appears to be some support for the use of IRT and chi square methods, somewhat less support for the use of transformed item difficulty and no justification for using discrimination differences.

The data for the comparison study of Rudner, Getson and Knight (1980) were produced with a Monte Carlo procedure specifying a priori the amount and type of differential item characteristics. Generation of the data consisted of specifying two groups, differentially performing items and examinee responses to these items. In order to relate characteristics of both items and examinees to responses, Birnbaum's (1968) three-parameter logistic model was used. The Getson, et al. study investigated the following seven techniques for identifying differentially performing items:

(1) TID-MA (transformed item difficulties - Major Axis). P values were transformed to deltas and absolute values of the perpendicular distance from the major axis indicated the magnitude of differential performance.

(2) TID-45 - P values were transformed to within-group z scores and the perpendicular item - 45° line distances indicated magnitude of differential performance.

(3) IRT-3 (three parameter logistic model)

(4) IRT-1F (Rasch model)

(5) IRT-1E (absolute differences in Rasch model parameters)

(6) CHI-5 (Chi-square with five total score intervals) and

(7) CHI-N (Chi-square with the number of possible score intervals minus number of cells with expected values less than 5).

These seven approaches were applied to seven tests of varying lengths (20, 30, 40, 50, 60, 70 and 80 items). Results indicated that none of the techniques were particularly affected by test length in terms of correlation between amount of differential performance generated and amount detected by each technique. This is surprising in light of the lower reliability associated with short tests. Hunter (1975) has shown that if reliability is low, observed item response functions are displaced from the true curves due to measurement error, and that the effect of such displacement is always to produce a spurious difference between observed item response functions. In addition the curve for the most able group always lies to the left and, therefore, above that of the least able group, which could be incorrectly interpreted as evidence of differential item characteristics.

Rudner, et al. found that when correlations between generated and detected differential characteristics were computed over all items, the IRT-3, CHI-5, and TID-45 methods were the most accurate, with respective correlations of .80, .73, and .68. The accuracy of the chi-square technique was reduced when more intervals were used. The authors conclude that the added cost of IRT-3 is offset by its increased accuracy.

Linn and Harnisch (1981) also discussed the advantages of invariant item parameters and the three-parameter model. They noted, however, that the three-parameter model requires very large sample sizes in order to obtain a large enough sample of minority groups for which the analysis is desired. They suggested a simpler model such as the Rasch one-parameter model as an alternative. However, they were concerned that group differences in difficulty estimates might be an artifact of differences in discrimination

or of the location of the lower asymptote. To resolve these problems, they developed an approach whereby the estimates of difficulty, discrimination and guessing of the three-parameter logistic model were obtained, using LOGIST, on all available cases in the sample. From these, P(θ) was obtained. The P(θ) estimates were compared to observed percent correct for the target group and for each of the subgroups. The difference was used as an index of the degree to which members of the groups respond to items in accordance with expected performance.

They applied this approach to the 46 multiple-choice mathematics items from the Illinois Inventory of Educational Progress administered to 2,055 eighth grade students, of whom only 283 were Black. For one item, the estimate of the a parameter was close to zero and that item was deleted from the analysis. The comparisons of D (the difference between observed and expected performance) were made separately for black and white students as a function of estimated values of θ. Students, divided by race, were further divided into quintiles within race on the basis of estimated θ's. P (the proportion in a subgroup expected to answer an item correctly according to the model), O (the observed proportion in the subgroup who answered the item correctly) and D (the difference between P and O) were computed for each item to determine whether there was a systematic difference between observed and expected performance at various θ levels. They also computed, within each quintile, average standardized difference scores (Z) as an overall index for each item.

They concluded that questions describing the metric system, definitions and graphs favored whites and that story problems involving money, unknown

symbol substitution and calculations were those on which black students did better than predicted by the model. In describing the test, they noted that it contained several types of items and believed that the differences in item types were sufficiently large that unidimensionality was, at best, only approximated. This problem, coupled with a sample of blacks that is too small for stable parameter estimation indicates that the conclusions drawn by the authors should be regarded with caution.

In their discussion, Linn and Harnisch (1981) list potential advantages and disadvantages of their approach. Indices are weighted by the distribution of estimated θ's in the target group, which may be an advantage since spurious differences may be indicated by comparison of item response functions in areas where sample size is low. For establishing minimal competency, the value of Z for particular regions of the θ scale may be helpful, and might also be used to develop a significance test. The disadvantages they noted, however, may outweigh the potential advantages. Item parameter estimates are influenced if the target group is part of the estimation sample, the values of D and Z depend on the definition and number of θ levels used to divide the subgroups, and D is sample-dependent. In addition, Hunter (1975) has shown that violations of unidimensionality may cause an artifactual detection of differential performance, and this set of items is clearly multidimensional.

Shepard, Camilli and Averill (1981), reviewed transformed item difficulties, item discriminations, chi-square and IRT approaches to the study of differential item characteristics from a conceptual standpoint and on the basis of technical soundness. They acknowledge that the three-

parameter logistic model is theoretically the most sound because of the sample invariance of the item parameters. However, because of the expense involved, the focus of their paper is on finding an acceptable substitute for the three-parameter model. Using an internal criterion, Shephard, et al. found a near perfect relationship ($r \geq .99$) between the transformed difficulty approach and the one-parameter model. They also found strong correlations between signed full chi-square and IRT-3 (three-parameter model) signed area indices. Correlations between IRT-3 and IRT-1 were much weaker. They conclude that a simpler method, such as chi-square, may approximate the three-parameter model well enough to recommend its use. However, this conclusion was based on samples of only 490 black, 551 chicano and 552 white students, despite the requirement of large samples (at least 1,000) for stable parameter estimation for IRT-3.

Sarrazin (1983) also compared transformed item difficulty, chi-square and IRT approaches to the study of differential item characteristics. To address the question of whether differential performance is a matter of group specificity, he administered a modified version of the intermediate level of the Canadian experimental version of the Otis-Lennon School Ability Test to 1,186 French and 601 English-speaking Quebec and Ontarian seventh and eighth graders matched on age, sex and grade level. The modified test consisted of 80 items from the American version increased by 20 items from the Otis- Lennon Mental Abilities Test. All items were translated into French, thereby introducing a potential source of error.

For the IRT analysis, separate analyses were conducted for each group. Sarrazin also did a principal component analysis on the b values. Since the number of students in the IRT analysis was relatively small and since more than 50 percent of the c parameter estimates didn't converge in one group, he set the c value to .75/k (where k is the number of choices) for all groups. In addition, he treated omitted responses as wrong answers. Factor analysis showed that the first factor accounted for approximately one-third of the total variance and he accepted this as meeting the minimum requirement for unidimensionality. Despite the departure from the requirement for a large enough sample for stable parameter estimation, Sarrazin concluded that previously found relationships among the methods did not hold, e.g., (1) the transformed item difficulty and IRT approaches agreed only in decisions based on item difficulty, (2) chi-square methods yielded comparable results, but didn't agree with results obtained with the transformed item difficulty approach and (3) the expected strong relationship between the three parameter model and chi-square approaches was absent. However, he states that independent of the method, items in the verbal comprehension category exhibited substantially more differential item characteristics than other item types. Sarrazin concludes that although the three-parameter model is usually favored, it is often difficult to find data which fits the model. As a solution, he offers a component score method which permits estimation of the standardized distance of each item to the principal axis and uses the percent of total variance explained by the second component to judge the importance of the differential item performance factor.

Shepard, Camilli and Williams (1984), in an effort to study problems of statistical artifacts in IRT differential item characteristic analyses, used data from the High School and Beyond data files of the National Center for Educational Statistics. randomly selected subsamples of 1,500 whites and 1,500 blacks for each of two black/white comparisons from the national probability sample of high school seniors and analyzed the senior mathematics and vocabulary tests. They factor-analyzed tetrachoric correlations using the total senior sample of 25,069 students. In both tests, the first factor accounted for 30 percent of the variance, providing reasonably strong evidence for unidimensionality. They also inspected plots of latent roots which showed a large distance between the first and subsequent eigenvalues.

In the application of IRT, LOGIST was used to estimate the item and person parameters. The c parameters were estimated in a combined analysis and then fixed at that value for the subgroups. The a's and b's were estimated separately for the groups and equated via a linear transformation of the b parameters, determined by a best fitting line that adjusted for the differences in average values. In computing means and variances, the inverse of the variance error in estimating b was used to weight b so that items with poorly estimated b's contributed least to the equating. The parameters for the black group were converted to the white scale. The θ values were also transformed, using the slope and intercept. The equating of the a parameters used the inverse of the slope determined for the b's.

For the purpose of quantifying differential item characteristics, they used both signed and unsigned indices. The unsigned indices were: (1) unsigned area between the two item response functions, (2) sum of squares 1: a self-weighting index in that scaled differences in probabilities are summed for every value of θ that occurs in the sample, (3) sum of squares 2: squared differences in probabilities weighted by the inverse of the variance error of the differences in item response functions for given θ's, and (4) chi-square. While these indices show the magnitude of the differences between item response functions, they do not reveal the direction of the difference. For this reason, the following signed indices were also used: (1) signed area: a negative sign was attached to the unsigned area if the item favored blacks in cases where the item response functions did not cross; otherwise, θ^* was found as the root of the equation and the signed area was the difference between the integral from -3 to θ^* and θ^* to +3, (2) sum of squares 3: analogous to sum of squares 1, but preserving the sign of the difference, (3) sum of squares 4: the weighted sum parallel to sum of squares 2, however, squared differences were weighted by the inverse of the variance error of the difference.

To assess the amount of artifactual differential item characteristics, in addition to the two black/white comparisons (i.e., W1B1=1,500 whites and 1,500 blacks and W2B2=1,500 whites and 1,500 blacks), analyses were conducted for randomly equivalent black and white groups, for extreme white groups (i.e., W1W2) and for randomly equivalent white groups.

For randomly equivalent black and white groups, a baseline for the magnitude of differential item characteristics obtained from the white/white

(W1W2) analysis was used. All values exceeding the largest number occurring in this white/white analysis were considered evidence of differential performance. They found 10 of the 29 math items consistently performed differentially across studies (3 in favor of blacks). For most identified items, the item response functions typically crossed within the θ region from -2 to +2, partially offsetting the performance differences in one region of the curve by reverse differences in the other.

A problem with signed indices was that they were large only if an item exhibited differential characteristics overall against a particular group. Shepard, et al. found that one item was artifactually identified. All of the indices (both signed and unsigned) were substantial in the first black/white comparison (W1B1), but not in the second (W2B2). The authors point out it was a difficult item for both groups and suggest that the a and b parameters were estimated in a region where there was little data, which is reflected in large standard errors. It was a clear outlier in the scattergram of the b's for the first comparison and sum of squares 2 and sum of squares 4 indices, which account for standard errors, had large values for this item. Inspection of the white/white comparisons showed smaller indices for all of the items than those obtained in the white/black comparisons. Even in this group, the artifactually identified item stands out.

The authors theorized that, even though the sample sizes were identical, restriction of range in the black group may have contributed to estimation problems and made the parameters more unstable. For this reason, they conducted a black/black analysis, which confirmed their hypothesis. They

suggest that despite the theoretical sample invariance of the IRT model, these models may be inadequate when there are large between-group differences in ability. They point out that equating procedures are more stable for horizontal than for vertical equatings in which the same test is administered to different grades, which is analogous to differential item characteristic studies where groups have large mean differences as in this study. In an effort to examine this further, a pseudo-black sample was selected from the white examinees, matched to the black group on the basis of the relative frequency distribution of the black math scores. For this comparison, there were very few large indices, suggesting that the differences in the white/black comparisons must be due to real differences in the way the items act across ethnic groups.

In addition, within-study (each ethnic group comparison, i.e., W1B1, W2B2, etc., constitutes a separate study) Spearman rank order correlation coefficients (i.e., intercorrelations of differential item characteristics indices) were obtained for each ethnic group comparison. Contrary to expectation, the correlations were not higher for those items identified as performing differentially. The signed indices were less highly correlated than the unsigned. Also the pattern of relationships was similar for both tests. Between-study comparisons showed how consistently an index ranked the items studied and supported the validity of the differential item characteristic indices. A correlation of .72 between sum of squares 2 statistics across studies when differential item characteristics were present was contrasted with the corresponding coefficients of .03 to .33 in the absence of differential performance. The authors recommend the sums of

squares 2, 3 and 4 indices as the most valid because they were found to be the most consistent in detecting differential item characteristics and had the lowest correlations when no differences were present.

Subkoviak, Mack, Ironson and Craig (1984) suggest that real data studies have the limitation that differential item characteristics are typically not manipulated and, therefore, it is difficult to determine which methods are detecting true differences. To address this problem, they constructed an instrument composed of 40 items from the verbal section of the College Qualification Test plus 10 items which tested black slang. One of the slang items was inserted at random among each 5 items. They administered this test to 1,008 black students enrolled in a small, predominantly black eastern university and 1,021 whites enrolled in a large, predominantly white midwestern university in order to insure that the slang items would favor the black group.

They used LOGIST to estimate the item response functions separately for each group and equated the resulting parameters. They computed the total area between the curves as an unsigned measure of differential item characteristics and attached a minus sign to area segments where the white curve was above the black, developing a signed index by summing across the segments. Although the authors evaluated several methods for detecting differential performance in addition to the three-parameter model (i.e., chi-square and transformed item difficulty), they found that the IRT model was the most effective at detecting a priori differences, with respective Pearson product-moment correlations of .872 and .875 for unsigned and signed indices. They concluded that the three-parameter model should be the method

of choice, provided sample size is large enough (1,000 or more per group), number of items is large (at least 40) and one has access to the necessary computer facilities.

In an effort to determine which techniques for detecting differential item characteristics best approximate the three-parameter model and which, therefore, may be used when the number of minority group examinees is too small to use the preferred IRT method, Shepard, Camilli and Williams (1985) compared chi-square, transformed item difficulty (TID) and pseudo-IRT indices with both real and simulated data. The IRT differential item characteristic indices used in their paper are described in the review of the Shepard, et al. (1984) study. The pseudo-IRT methodology evaluated was that proposed by Linn and Harnisch (1981). Their real data consisted of a random subsample of 1,000 whites and 300 blacks selected from the sample used in Shephard, et al. (1984). Their instrument consisted of the 32-item Mathematics test, described in Shephard, et al.(1984). In that study, they found that verbally-presented math items consistently performed differentially with respect to blacks and whites.

In the Shepard et al. (1985) research, the interest was in the results of the replicated black-white comparisons. Randomly-equivalent groups of blacks and whites were used in two applications of IRT differential item characteristic procedures. The magnitude of the differential performance indices was interpreted in light of the two white groups' performance. Cross-validation of the comparisons resulted in the identification of 10 items (7 of which performed differentially against blacks and 3 against whites), which became the criterion of detection in the 1985 study.

Shepard, et al. (1985) indicate that this standard of comparison is acceptable without concern about estimation errors, due to replication of IRT results.

In addition to analyses on the previously described sample, distributions were matched (on the basis of the total score frequency distribution for blacks) and chi-square and delta analyses were conducted on these matched distributions.

The second part of the study consisted of the application of the methodologies of interest to two sets of generated data, one containing 54 items simulated to be free of differential item characteristics, the other also containing 54 items, the last 18 of which were simulated to perform differentially (9 weakly and 9 moderately). For this part of the study, a baseline was established and matched distributions were also created.

Results for Part I indicated that the pseudo-IRT method correlates highest with the sum of squares 4 index in the Shepard et al. (1984) study, followed by chi-square and TID. There was a considerable improvement in the correlations of TID with the criterion when matched groups were used. The pseudo-IRT approach identified 6 of the 10 differentially performing items and made one false positive error, i.e., there is 83% agreement. Chi-square appeared to be equally good in terms of percent of agreement. Even with matched distributions, TID did not do as well as the other methods.

The same pattern is to be found in Part II. When compared to the full IRT analysis for this sample size, the pseudo-IRT approach performed equally well in terms of detecting moderately differentially performing items. However, none of the approximation techniques were very accurate in detecting weak amounts of differential performance.

Petersen (1977) agrees that using IRT to study differential item characteristics is the only known sound theoretical approach, but states that IRT does have its limitations. She lists the limitations as being that it assumes the average item is free of differential characteristics, is incapable of detecting consistent differential performance across items due to confounding with group differences in mean ability, and gives no information with respect to why an item may be performing differentially.

The foregoing literature review has revealed that some limitations of the IRT approach appear consistent across studies, such as the frequent non-convergence of c parameters. In addition, there is some question regarding the appropriateness of the significance tests to compare item response functions across groups. Lord (1980) notes that the tests are asymptotic, assume that θ's are known and only apply to maximum likelihood estimates. An additional source of concern is the fact that in the absence of an external criterion, the total test is used to determine estimated θ's, and yet, the assumption that the test as a whole is free of differentially performing characteristics may not be a tenable one.

Further, the signed and unsigned indices of differential item characteristics, according to Ironson (1982) have inherent problems in that the c parameters are fixed, the equating line is not completely satisfactory and the ability estimate may be based on differentially performing items. Ironson (1982) points out that the distribution of ability may be such in a particular group that, over the area containing the most subjects in the comparison group, it may be possible to only poorly estimate item performance. She also notes that the same predictions may be made by items

with quite different parameters, i.e., there may be no practical difference in the item response functions despite the difference in parameters. There is little information available regarding how the significance test for equality of parameters and the area measure indices relate and they would not necessarily agree in the items they indicate as performing differentially.

In summary, there are several advantages inherent in the application of the three-parameter model to the study of differential item characteristics, the most notable being:

- o the relationship of the probability of a correct response to the underlying ability level of the examinee and the item characteristics,
- o the ability of the model to take guessing into account, and
- o the fact that item parameters are independent of the sample and, therefore, theoretically invariant across groups.

In terms of the previously discussed limitations of analysis of variance, transformed item difficulty and chi-square approaches (i.e., their sample dependence and confounding of ability with characteristics of the items), it should be clear that IRT is the theoretically preferred approach to the study of differential item characteristics. It should also be noted that the three parameter model is preferred to the Rasch model because of the information contained in the c parameter.

This endorsement of the three-parameter model is, however, not without qualification. The limitations of the model are such that it is often difficult to use and in many cases impossible to apply.

The most salient limitations are summarized as follows:

- o It is expensive.
- o Sample sizes must be large (at least 1,000) and there should be at least 40 items to be analyzed.
- o The c parameters often fail to converge.
- o The appropriateness of asymptotic significance tests is questionable.
- o The test must be unidimensional. This constraint, however, is true of all methods for detecting differential item characteristics.

For tests which lack the requisite number of items and/or for samples too small to meet the criterion, the literature suggests that pseudo-IRT and chi-square approaches may be suitable substitutes for the three-parameter model. Clearly, tests of model fit are imperative to rule out multidimensionality, which would engender artifactual results.

It is noted, in conclusion, that further research is needed in the field of IRT in relation to the study of differential item characteristics to resolve conflicts of agreement between significance tests for equality of parameters and area measure indices in terms of the items they indicate as performing differentially and to develop theoretically sound significance tests. Studies comparing IRT to other methodologies recently applied to the study of differential item characteristics, such as the standardization approach of Dorans and Kulick (1983), the Mantel-Haenszel approach endorsed by P. Holland (personal communication, Sept. 1985) and full log-linear models are also needed to determine the most acceptable approximation techniques.

REFERENCES

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

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

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

Dorans, N. J. (1979). The Need for a Common Metric in Item Bias Studies. Research Report 79-20, U.S. Office of Personnel Management, Washington, D.C.

Dorans, N. J. & Kulick, E. (1983). Assessing unexpected differential item performance of female candidates on SAT and TSWE forms administered in December 1977: An application of the standardization approach. ETS Research Report (RR-83-9). Princeton, NJ: Educational Testing Service.

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

Haebara, T. (1979). A Method for Investigating Item Bias Using Birnbaum's Three-Parameter Logistic Model. Iowa Testing Programs Occasional Papers, Number 25. Iowa City: Iowa Testing Programs, University of Iowa.

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

Intasuan, P. (1979). A comparison of three approaches for determining item bias in cross-national tests. Unpublished doctoral dissertation, University of Pittsburg.

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, Johns Hopkins University Press, Balt., Md., Ch. 5, 117-160.

Ironson, G. H. & Subkoviak, M. J. (1979). A comparison of several methods of assessing item bias. Journal of Educational Measurement. 16, 4, 209-225.

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

Lord, F. M. (1952). A theory of test scores. Psychometric monograph, No. 7. Chicago: University of Chicago Press.

Lord, F. M. (1977). A study of item bias, using item characteristic curve theory. In Poortinga, Y.H. (Ed.), Basic problems in cross-cultural psychology, Amsterdam: Swets and Zeitlinger, 19-29.

Lord, F. M. (1980). Applications of Item Response Theory to Practical Testing Problems. Lawrence Erlbaum Assoc., Hillsdale, N.J., pp. 212-223.

Petersen, N. S. (1977). Bias in the selection rule -- bias in the test. Paper presented at the Third International Symposium on Educational Testing, University of Leyden, The Netherlands.

Pine, S. M. (1977). Applications of item response theory to the problem of test bias. NR No. 150-383, Personnel and Training Research Programs, Office of Naval Research.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen: Nielson and Lydiche (for Denmarks Paedagogiske Institute.

Rudner, L. M. (1977). An approach to biased item identification using latent trait measurement theory. Paper presented at the Annual Meeting of AERA, New York, NY.

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

Sarrazin, G. (1983). The detection of item bias for differential cultural groups using latent trait and chi-square methods. Paper presented at the Joint European Meeting of Psychometric and Classification Societies, Jouy-en-Josas, France.

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, 4, 317-375.

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

Scheuneman, J. (1979). A new method of assessing bias in test items. Journal of Educational Measurement, 17, 1-10.

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.

Wood, R. L., Wingersky, M. S. & Lord, F. M. (1976). LOGIST - A computer program for estimating examinee ability and item characteristics curve parameters. Research Memorandum 76-6. Princeton, NJ: Educational Testing Service.

Wright, B. D., Mead, R. & Draba, R. (1976). Detecting and correcting test item bias with a logistic response model. Research Memorandum Number 22, Statistical Laboratory, Dept. of Education, University of Chicago.