DOCUMENT RESUME

ED 380 499 TM 022 862

AUTHOR Dorans, Neil J.; Potenza, Maria T.
TITLE Equity Assessment for Polytomously Scored Items: A Taxonomy of Procedures for Assessing Differential Item Functioning. Research Report RR-94-49.
INSTITUTION Educational Testing Service, Princeton, N.J.
PUB DATE Oct 94
NOTE 36p.; Version of a paper presented at the Annual Meeting of the National Council on Measurement in Education (Atlanta, GA, April 11-15, 1993).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Classification; *Educational Assessment; Educational Change; *Equal Education; *Item Bias; Multiple Choice Tests; Scores; *Test Items
IDENTIFIERS *Binary Scores; *Polytomous Scoring; Reform Efforts

ABSTRACT

Educational reform efforts have led to increased use of alternatives to the traditional binary-scored multiple choice item. Many stimuli employed by these alternative assessments yield complex responses that require complex scoring rules. Some of these new item types can be polytomously-scored. Differential item functioning (DIF) assessment is a form of equity assessment that attempts to identify items for which subpopulations of examinees exhibit performance differentials that are inconsistent with the performance differentials typically seen for those subpopulations on collections of items that purport to measure a common construct. Any DIF technique should be evaluated in terms of how well it meets certain statistical and practical criteria before it can be concluded that the items associated with alternative forms of assessment can be adequately tested for DIF. DIF methodology is well-defined for traditional, binary-scored multiple-choice items. This paper provides a classification scheme of DIF procedures for binary-scored items that is applicable to new DIF procedures for polytomously scored items. In the process, a formal development of a polytomous version of a binary DIF technique is presented. Finally, several polytomous DIF techniques are evaluated in terms of statistical and practical criteria. (Contains 65 references and 3 tables). (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

RESEARCH REPORT

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it
☐ Minor changes have been made to improve reproduction quality
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
R. Coley
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

EQUITY ASSESSMENT FOR POLYTOMOUSLY SCORED ITEMS: A TAXONOMY OF PROCEDURES FOR ASSESSING DIFFERENTIAL ITEM FUNCTIONING

Neil J. Dorans
Maria T. Potenza

ED 380 499

Educational Testing Service
Princeton, New Jersey
October 1994

TM022862

Equity Assessment for Polytomously Scored Items: A Taxonomy of Procedures for Assessing Differential Item Functioning[1]

Neil J. Dorans and Maria T. Potenza
Educational Testing Service

[1]The authors wish to thank Rebecca Zwick, David Thissen, Barbara S. Plake, Eiji Muraki, Roger E. Millsap, Howard T. Everson, Fritz Drasgow and Hua-Hua Chang for their insightful comments on an earlier version of this paper. A previous version of this paper was presented at the 1993 annual meeting of the National Council on Measurement in Education, Atlanta, GA. Partial funding for this paper was provided by the College Board Division of the Educational Testing Service.

Copyright © 1994. Educational Testing Service. All rights reserved.

Abstract

Educational reform efforts have led to increased use of alternatives to the traditional binary–scored multiple choice item. Many of the stimuli employed by these alternative assessments yield complex responses that require complex scoring rules. Some of these new item types can be polytomously–scored. Differential item functioning (DIF) assessment is a form of equity assessment. DIF assessment attempts to identify items for which subpopulations of examinees exhibit performance differentials that are inconsistent with the performance differentials typically seen for those subpopulations on collections of items that purport to measure a common construct. Any DIF technique can be evaluated in terms of how well it meets certain statistical and practical criteria. These DIF assessment criteria need to be attended to before we can conclude that the items associated with alternative forms of assessment can be adequately tested for DIF. DIF methodology is well–defined for traditional, binary–scored multiple–choice items. This paper provides a classification scheme of DIF procedures for binary–scored items that is applicable to new DIF procedures for polytomously–scored items. In the process, a formal development of a polytomous version of a binary DIF technique is presented. Finally, several polytomous DIF techniques are evaluated in terms of statistical and practical criteria.

Equity Assessment for Polytomously Scored Items: A Taxonomy of Procedures for Assessing Differential Item Functioning

Over the past decade, the clamor over the poor quality of education in the United States has led to an outcry for educational reform (National Commission on Excellence in Education, 1983; U.S. Congress, 1992). Educational testing is increasingly viewed as one of the primary tools for implementing educational reform in the United States (e.g. America 2000, National Assessment for Educational Progress). Furthermore, tests have come to be viewed as a means of ensuring greater accountability in the educational process (Madaus, 1985). This widespread emphasis accorded to educational testing has given rise to the idea that tests can lead to changes in curriculum and instruction. Because tests are thought to play such a powerful role in influencing the learning process, many are advocating the redesign of tests to support educational goals (Morrison, 1992). Politicians and educators are arguing for "new and better" assessment methods—methods which more closely resemble what goes on in classrooms (National Educational Goals Panel, 1991). Test reformers are calling for assessment procedures which consist of "authentic" tasks that students should practice. This has resulted in the proliferation of "performance" or "authentic" classroom assessment procedures as replacements for traditional tests (see Bennett and Ward, 1993 for an in-depth examination of assessment methodologies arising from the authentic assessment movement). Thus, most proposals for national testing programs or systems of assessment do not call for the implementation of "known" testing technology, i.e., the multiple-choice, norm-referenced achievement test (Morrison, 1992). To the contrary, the multiple-choice test, with its sound psychometric properties developed over nearly a century, is viewed increasingly as a task irrelevant to the learning and development of students.

Performance assessment includes a broad range of testing methods that often require students to create an answer rather than select a response. Performance assessment item types span a continuum from multiple-choice to presentation/performance according to the degree of constraint placed on the examinee's response (Bennett & Ward, 1993). At the upper end of the continuum, performance assessments could require students to write an essay, carry out experiments, create and defend a position in an oral

performance, or to assemble a "portfolio" over a period of time to illustrate growth in a particular skill or domain (Camp, 1993; Morrison, 1992; Valencia & Calfee, 1991).

Assessment methods such as these are complex, both in terms of the "item" or task stimuli and the type of response which the stimuli produce. Such assessments require new methods for measuring student responses (i.e., scoring). Often these assessment techniques are scored in a "polytomous" manner, that is, they use scoring rubrics that have several categories and assume an inherent "order" of degree of correctness of the response. This paper will focus on the polytomously–scored item type.

Equity Concerns

Differential item functioning (DIF) refers to a psychometric difference in the way an item functions for two groups. DIF indicates a difference in item performance between two comparable groups of examinees, that is, groups that are matched with respect to the construct being measured by the test. The comparison of matched or comparable groups is critical because it is important to distinguish between differences in item functioning and differences in group ability. The vast majority of multiple choice tests are rights–scored (i.e. each item is scored either as right or wrong). Even when a multiple–choice test is not rights–scored, it is often analyzed as if it were (Dorans, 1991). Most procedures utilized to assess DIF presume that items are scored in this binary fashion (Holland & Wainer, 1993). Currently there are numerous methods for conducting DIF assessment for binary–scored items (see Millsap & Everson, 1993; Scheuneman & Bleistein, 1989 for a review).

Educational reform efforts have led to increased use of alternatives to the traditional binary–scored multiple choice item. Many of the stimuli employed by these alternative assessments yield complex responses that require complex scoring rules. Some of these new item types can be polytomously–scored. Recently, several procedures have been proposed for the assessment of DIF for polytomously–scored items (Chang, Mazzeo, & Roussos, 1993; Grima, 1993; Muraki, 1993; Rogers & Swaminathan, 1993; Welch & Hoover, 1993, Wilson, Spray, & Miller, 1993; Zwick, Donoghue, & Grima, 1993b). However, several important methodological issues will need to be addressed in the transition from binary to polytomous items. These issues can be subdivided into two classes: (1) issues pertaining to the validity

of the rules for assigning scores to stimuli and the quality of the matching variable, and (2) issues directly related to the statistical and practical utility of the particular DIF procedure. A meaningful DIF study requires satisfactory resolution of the first class of issues. The second class contains criteria for evaluating alternative DIF procedures.

The goal of this paper is threefold. First, we suggest a classification scheme for DIF procedures used with binary-scored items, and then apply this classification system to DIF procedures for polytomously-scored items. Second, we delineate several issues associated with the extension of current DIF procedures to performance assessments in which polytomous scoring rules are utilized. Finally, we propose criteria for the evaluation of polytomous DIF techniques and evaluate a selected set of polytomous DIF techniques in terms of these criteria.

Framework for the Classification of DIF Procedures

Two classes of DIF procedures exist for binary items: observed-score approaches and latent-variable approaches (Millsap & Everson, 1993). Both classes assume that the items studied for DIF measure the same dimension as the matching variable, i.e. they presume unidimensionality. The fundamental difference between these two classes of approaches is that the former uses an observed-score as the matching variable, while the latter uses an estimate of latent ability, which is a function of observed data. This distinction has implications for how DIF is defined and measured. In addition to this distinction by type of matching variable, we distinguish between procedures that employ a functional form for the relationship between item score and the matching variable (i.e. parametric procedures) and those that do not (i.e. nonparametric procedures). Other classification schemes impose a dichotomy similar to the distinction between observed score and latent variable approaches (Scheuneman & Bleistein, 1989; Wainer, 1993). However, these schemes do not make a clear distinction between the type of DIF being assessed and the amount of structure imposed on the data by the technique. That is, they omit the important distinction between procedures which define a functional form for the item score/matching variable relationship and those that do not. Consequently, the false impression may be conveyed that all latent variable models employ a parametric form, while all observed score approaches do not. The framework

we present adds this important distinction and can be used to classify both binary and polytomous DIF procedures. Parametric approaches to DIF detection require the assumption that the model for describing the relationship between item performance and the matching variable is correctly specified.

A problem associated with the parametric approach is that the detected DIF is often an artifact of model mispecification. In addition, very large sampling covariation among parameter estimates is often a problem for parametric approaches that employ several parameters (Lord, 1980; Ramsay, 1991; Thissen & Wainer, 1982). While the nonparametric procedures are relatively free of model mispecification and collinearity problems, they require sufficient data to directly estimate the item/test regressions. In small samples, these procedures may produce unstable results due to the effects of sampling error.

These two definitional distinctions, matching on an observable vs. matching on a model-based estimate of an unobservable or latent variable, and whether the approach posits a parametric form for the relationship between item score and the matching variable can be crossed to produce Table 1 for binary DIF procedures.

Observed-Score DIF Procedures

Although there are many observed-score procedures for assessing DIF on binary-scored items (Holland & Wainer, 1993; Scheuneman & Bleistein, 1989: Shepard, Camilli, & Williams, 1985), we will focus on three methods because initial attempts have been made to extend these procedures to the case of polytomous DIF: the standardization (STND) procedure (Dorans & Kulick, 1983; 1986), the Mantel-Haenszel (MH) method (Holland & Thayer, 1988), and a logistic regression (LRDIF) approach (Swaminathan & Rogers, 1990). Each of these procedures are observed-score approaches because they share a common definition of null-DIF at the item level: "An item is unbiased if, for all individuals having the same score on a homogeneous subtest containing the item, the proportion of individuals getting the item correct is the same for each population group being considered" (Scheuneman, 1975, p. 2). All three procedures employ an observed score measure of the construct of interest as a matching variable. Hence, they state

that there is no differential item functioning between groups after they have been matched on an observed score, usually the total score. None of these three observed–score methods postulate a psychometric or a cognitive model of item or test performance (see Dorans & Holland, 1993; Swaminathan & Rogers, 1990 for a more complete description of these observed–score DIF procedures).

Table 1

Cross–Classification of Binary DIF Procedures.

	A Parametric Form for Relationship Between Item Score and the Matching Variable	**No Parametric Form** for Relationship Between Item Score and the Matching Variable
Observed–Score Matching Variable	Logistic Regression	Mantel–Haenszel
		Standardization
Latent–Variable Matching Variable	General IRT–LR	SIBTEST
	Limited Information IRT–LR	
	Log linear IRT–LR	
	IRT D^2	
	Lord's Chi Square	

Binary Non–Parametric Observed–Score DIF Procedures

STND and MH are both observed–score approaches for binary DIF that do not specify a parametric form for the relationship between item–scores and the matching variable (Dorans & Holland, 1993).

Standardization (STND). The null–DIF definition for the STND method states that at each level of the matching variable there is no difference in proportions correct between the focal group (the focus of the DIF analysis) and the reference group (the basis for comparison). This can be conceived of as zero difference in expected item score given the matching variable, or as no difference between empirical item test regressions for the

focal and reference groups. This approach does not use any parametric function to fit either the empirical item test regressions or the difference between empirical item test regressions of the focal and reference groups.

An average overall index of DIF, which is referred to as STD P–DIF, is obtained by averaging differences in expected item scores across levels of the matching variable, weighting each difference by focal group relative frequencies. A standard error has been developed to quantify the stability of this index (Dorans & Holland, 1993), and it has been shown to perform well in practical applications (Donoghue, Holland, & Thayer, 1993). No formal statistical test of the null hypothesis has been developed for the standardization approach, although a test statistic involving the ratio of STD P–DIF to its standard error can be employed.

Mantel-Haenszel (MH). The null–DIF definition for the MH method is that the odds for responding correctly is the same in both the focal group and the reference group, given a level of the matching variable across all M levels of the matching variable (i.e., total score). This definition has been shown to be equivalent to STND's definition of null–DIF, which is in terms of proportions correct (Dorans & Holland, 1993).

The MH approach is sometimes viewed as parametric because it postulates a particular statistical model, known as the constant odds–ratio model, as a particular type of violation of null–DIF. In other words, the MH approach measures amount of DIF under the restriction that the odds–ratio is the same or constant across all score levels. Hence, it is often referred to as a uniform DIF model. It does not, however, postulate a particular parametric form for the odds, for either the focal or reference group, as a function of the matching variable.

Mantel and Haenszel (1959) provided an estimate of the constant odds–ratio (α_{MH}) that ranges from 0 to ∞ with a value of 1 indicating null–DIF. In general, odds are converted to log odds because the latter is symmetric around zero and easier to interpret. Holland and Thayer (1985) converted α_{MH} into a difference in deltas, MH D–DIF, via a log odds transformation. The standard error of the MH D–DIF has been studied extensively and performs well in DIF contexts (Donoghue, Holland & Thayer, 1993). In addition, the MH procedure is associated with a well–established significance test (Mantel & Haenszel, 1959).

Polytomous Non–Parametric Observed–Score DIF Procedures

STND and MH are two closely related binary DIF techniques that measure DIF identically when generalized to the polytomous situation. One generalization of the MH procedure is the Mantel (1963) procedure. Two mathematically equivalent measures of polytomous DIF have been suggested as useful supplements to the hypothesis test statistic for the Mantel procedure (Dorans & Schmitt, 1993; Zwick, Donoghue & Grima, 1993b). Because the extension of the STND model to polytomously–scored items has not been developed in mathematical detail elsewhere, we will do so here, and point out its relationship to the polytomous DIF version of the Mantel procedure.

Polytomous STND. The general STND approach involves a comparison of two empirical item–test regressions, in which differences in these regressions at each score level are weighted by the relative frequencies of focal group members at that score level. These weighted differences are then summed across score levels to arrive at a measure of DIF. The distinction between binary DIF and polytomous DIF is the number of levels of the dependent variable, i.e., the item score. For binary items, the STD P–DIF index is an average weighted difference in proportions correct (the expected item score under binary scoring) across score levels. The more general index is STD ES–DIF, or standardized expected item score DIF. For the general case, we assume that there is: (1) a matching variable, X, with M levels, m = 1, M; (2) an ordered item score, Y, with K levels, k = 1, K; and (3) two groups: r (reference) and f (focal).

The polytomous version of STND starts with the computation of expected item scores for both the focal group, $E_{fm}(Y|X)$, and the reference group, $E_{rm}(Y|X)$, via

$$E_{fm}(Y|X) = \sum_k N_{fmk}\, Y_k / N_{fm}, \qquad (1)$$

and

$$E_{rm}(Y|X) = \sum_k N_{rmk}\, Y_k / N_{rm}, \qquad (2)$$

where N_{fmk} is the number of examinees in the focal group at score level m with item score Y_k, and N_{fm} is the total number of examinees in the focal

group at score level m. The terms N_{rmk} and N_{rm} are parallel reference group frequencies. The item score variable, Y_k, can take on any ordered values , including 1, 2, 3, ... K.

As with binary STND, the next step is to take differences in expected item scores at each level of the matching variable,

$$D_m = E_{fm}(Y \mid X) - E_{rm}(Y \mid X), \quad (3)$$

and weight these differences by focal group relative frequencies (Dorans & Kulick, 1986), to obtain

$$\text{STD ES-DIF} = \sum_m N_{fm} D_m / N_f \quad (4)$$

where, N_f is the total number of focal group examinees.

In the adaptation of the Mantel procedure to polytomous DIF, the expression in Equation 4 is defined as the standardized mean difference (Zwick et al., 1993b). In addition, there is a test statistic associated with the Mantel approach,

$$MNTL = (\sum_m F_m - \sum_m E\{F_m\})^2 / \sum_m VAR_m\{F_m\}, \quad (5)$$

where

$$F_m = E_{fm}(Y \mid X) N_{fm}. \quad (6)$$

and $E\{F_m\}$ and $VAR\{F_m\}$ are the mean and variance of F_m under the hypothesis of no association between group and item score given the value of the matching variable. Under the no association null hypothesis, MNTL is distributed as a chi-square with one degree of freedom (Mantel, 1963; Zwick et al, 1993b)

HW1 and HW3 Approaches. Recently, another pair of test statistics have been proposed for detecting departures from null-DIF for polytomously-scored items (Welch & Hoover, 1993). Both indices can be described using a general standardization framework in which differences in expected item scores are weighted across levels of the matching variable to

arrive at a summary measure of DIF. One index, HW1, takes the difference in expected item performance at each level of the matching variable and converts it into a t-statistic by dividing by a pooled standard error of the mean difference. These t-statistics are then summed across levels of the matching variable, and divided by the square root of the sum of the variances of these independent t-statistics. The resultant statistic is normally distributed with a mean of 0 and a standard deviation of 1. The second index, HW3, weights each test statistic by the reciprocal of its sampling variance. A correction factor is employed at each level of the matching variable to correct for bias in small samples. The resultant statistic is normally distributed with a mean of 0 and a standard deviation of 1. Thus, both HW1 and HW3 fall within the general standardization framework in which differences in expected item scores are averaged across levels of the matching variable using weights that are driven by statistical considerations. However, HW1 and HW3 are test statistics; their magnitudes are sample–size dependent, so they are not measures of the amount of DIF.

Generalized Mantel-Haenszel (GMH) Approach. The generalized MH procedure is another generalization of the binary MH procedure (Mantel and Haenszel, 1959). Whereas the polytomous STND procedure and the Mantel procedure emphasize expected (average) item scores when comparing focal and reference groups, the generalized MH (GMH) procedure compares entire item response distributions, conditioned on the matching variable.

The test statistic for the Mantel procedure is univariate, for the weighted linear composite of the item scores that defines the expected score. The test statistic for the GMH is multivariate normal and distributed with K–1 degrees of freedom under the null hypothesis of no association between item responses and group, given a fixed value of the matching variable (Zwick et al, 1993b). This test statistic is sensitive to any differences in conditional response patterns between the focal and references groups, while the Mantel and polytomous STND approaches are sensitive to differences between the means of these conditional distributions.

An interpretable overall measure of amount of polytomous DIF is difficult to develop, but a series of partial odds ratios can be used to describe the amount of DIF (Zwick et al, 1993b). There are many collections of partial odds–ratios, however, just as there are many sets of contrasts available in an ANOVA.

Binary Parametric Observed–Score DIF Procedure

Logistic Regression (LRDIF) Approach. The LRDIF approach is an observed–score method that specifies a particular parametric form for the item score/matching variable relationship. Swaminathan and Rogers (1990) postulate a statistical model, logistic regression, for the probability of answering an item correctly for a fixed observed score. Their definition of null–DIF is a variation on the more generic STND definition, because they postulate a parametric functional form for the empirical regression employed by STND. Significance tests exist for both uniform DIF and cross–over DIF, i.e. item–test regressions with intersection points.

The MH procedure can be viewed as a special case of the general logistic regression model in which the matching variable is discrete, as is often the case, and the interaction term between score level and group equals zero (Swaminathan & Rogers, 1990). Thus, the LRDIF technique shares the definition of null–DIF used in both the MH and STND approaches, specifically that there is no differential item functioning between groups after they have been matched on an observed score measure of the construct of interest.

Descriptive measures of an item's degree of DIF are essential to DIF assessment. Both the MH and STND procedures have measures, MH D–DIF and STD P–DIF, respectively, that are meaningfully defined (Dorans & Holland, 1993) and well–studied (Allen & Holland, 1993; Donoghue, Holland, & Thayer, 1993; Longford, Holland, & Thayer, 1993). Swaminathan and Rogers (1990) do not propose a descriptive statistic for degree of DIF for the LRDIF technique.

Polytomous Parametric Observed–Score Procedure

Polytomous LRDIF. The logistic regression DIF procedure can be extended to the polytomous case (Miller & Spray, 1993; Rogers & Swaminathan, 1993). Like the GMH approach, polytomous LRDIF can be used in many ways to analyze the data. Each approach involves a different set of pairwise comparisons between score categories or combinations of score categories. One approach is to compare item performance in adjacent categories across groups . This requires fitting K–1 logistic regression models and involves 2(K–1) significance tests, where K is the number of levels of the polytomous score. Continuation–ratio logits and the proportional odds

model are two other polytomous LRDIF approaches that produce different sets of K–1 logistic regression functions (Agresti, 1990). The absence of a descriptive measure of DIF, in conjunction with the need to examine the K–1 logistic regression functions, makes the polytomous LRDIF procedure difficult to interpret and at times "unwieldy" (Miller & Spray, 1993). A further complication is that the results obtained may differ across models, because each estimates a different sets of odds ratios.

Latent–Variable DIF Procedures

A second class of DIF techniques for binary items is rooted in strong true score theory (e.g. item response theory (IRT)) or weak true score theory (classical test theory; Lord & Novick, 1968). Central to these psychometric models is the decomposition of observed test performance into a reliable portion and an unreliable portion. The reliable portion is often referred to as latent ability, underlying proficiency, or true score. A fundamental difference between the latent–variable approaches and the observed–score approaches is the utilization of estimates of the latent ability or true score instead of observed score as either an implicit or explicit matching variable. As with the observed–score approaches, the latent–variable methods can be divided on the basis of whether or not they specify a parametric form for the item response function.

Binary Parametric Latent-Variable DIF Procedures

There are several variations on one theme among parametric item response theory approaches. These variants state that an item has DIF if "...an item has a different item response function for one group than for another..." (Lord, 1980, p. 212). A variety of parametric IRT DIF procedures exist. They differ with respect to the particular parameterization of the item response function (IRF) assumed, the type of parameter estimation employed, and the types of significance tests used to assess differences in item parameters (see Thissen, Steinberg & Wainer, 1993 for a detailed description of these approaches).

The most general approach is the General IRT–Likelihood Ratio (LR) approach (Thissen, Steinberg & Wainer, 1988), which uses the Bock–Aitken (Bock & Aitken, 1981) marginal maximum likelihood estimation algorithm

to estimate parameters for a wide variety of models. A second approach, Log-Linear IRT-LR employs maximum likelihood estimation (Kelderman, 1989). A third approach, Limited-Information IRT-LR employs normal ogive IRT models with generalized least squares estimation of parameters (Muthen & Lehman, 1985). Each of these three approaches employ likelihood ratio (LR) tests to assess the significance of DIF effects, contrasting a compact model in which focal and reference group IRFs are identical with an augmented model in which the IRFs differ.

These three likelihood ratio approaches have been previously evaluated (Thissen, Steinberg & Wainer, 1993). The least applicable for DIF analysis for traditional binary multiple-choice items is the Log-Linear IRT-LR procedure because it is restricted to classes of Rasch models that do not permit items to have different discrimination or non-zero asymptote parameters. The normal ogive IRT models employed by the Limited-Information IRT-LR approach similarly do not permit non-zero asymptotes, and require larger sample sizes than the other LR methods, but they can be used to test for DIF within a multidimensional model. Since the General IRT-LR approach accommodates a wide variety of IRT models, it is the approach that is least likely to confound DIF between the focal and reference groups with lack of fit of the IRT model to the data. However, each of the three LR approaches can be labor and computationally expensive, especially the General IRT-LR approach, because each item is studied separately, and two sets of item parameter estimates (or more) are required for each item. A standardized DIF statistic for any IRT DIF model has been proposed (Wainer, 1993); it is based on the focal group weighting procedure from the STND approach (Dorans & Kulick, 1986).

A fourth IRT-based approach for DIF assessment analyzes all of the items simultaneously. The IRT-D^2 approach uses the Bock-Aikten marginal maximum likelihood EM algorithm followed by one or two iterations of the Bock and Lieberman (1970) direct Newton-Raphson algorithm to estimate item parameters in the reference and focal groups (Bock, Muraki, & Pfeiffenberger, 1988). The Newton-Raphson algorithm provides standard errors for the item parameter estimates. The three-parameter logistic model used but only the difficulty parameters differ between groups. Unlike the LR procedures, the IRT-D^2 approach uses the ratios of parameter differences to their standard errors to evaluate the significance of observed differences. One

descriptive index of DIF which can be used is the standardized index of bias (Muraki & Engelhard, 1989). An alternative index to measure amount of DIF (in the latent variable metric), is the difference between item difficulty estimates. This approach is analogous to the MH delta difference. A standardized DIF statistic using focal group weighting could also be employed.

Lord (1980) also suggested a procedure within this category, a procedure that has come to be known as Lord's chi-square approach (McLaughlin & Drasgow, 1987). As with the IRT-D^2 method, Lord's chi-square approach presumes that the three parameter logistic model fits the data in both the focal and the reference group. Both discrimination and difficulty, however, are allowed to differ between groups. A chi-square test is used to simultaneously test the null hypothesis of no differences in both parameters across groups.

Polytomous Parametric Latent-Variable DIF Procedures

There are several IRT–based models that can be used with polytomously–scored items. Some of these models posit a parametric form for the probability of choosing each category as a function of underlying proficiency. These parametric models fall into two general classes: "difference" models and "divide–by–total" models (Thissen & Steinberg, 1986). For the difference model class, the parametric form for the probability of choosing category k, P(k), is written most simply as a difference between two adjacent cumulative probabilities P*(k) – P*(k+1), where P*(k) is the probability of a response in category k and above. The graded response model is an exemplar for this class (Samejima, 1969). For the divide–by–total class of models, the parametric form for the probability of choosing category k is written most simply as an exponential divided by a sum of exponentials (Thissen & Steinberg, 1986). The nominal response model (Bock, 1972), the partial–credit model (Masters, 1982), and the rating–scale model (Andrich, 1978) are all examples of the "divide–by–total" class of models. The multiple choice model is a modified version of the nominal model that allows for non–zero lower asymptotes in the expressions for P(k) (Thissen & Steinberg, 1984).

The general nominal response model has been adapted for the study of DIF on item sets (Wainer, Sireci, & Thissen, 1991) and the same methodology can be employed to study polytomous DIF. The approach employs a series of

likelihood ratio (LR) tests to assess the significance of DIF effects, contrasting a compact model in which focal and reference group item category response functions are identical with different augmented models in which the item category response functions differ.

In the partial–credit model (Masters, 1982), the propensity to select category k on item i is expressed as

$$P_{ik} = \exp\{\sum_{v=1}^{k} (\theta - b_{iv})\} \; / \; \sum_{c=1}^{K} \exp\{\sum_{v=1}^{c} (\theta - b_{iv})\}, \qquad (7)$$

where the b_{iv} are the points of intersections for adjacent categorical response curves, called step parameters, and θ is the individual's ability or proficiency.

In this divide–by–total model, all items have the same discrimination parameter, which equals 1.0 and does not appear in Equation 7. The rating–scale model (Andrich, 1978) can be derived from this model by decomposing

$$b_{ik} = b_i - d_k, \qquad (8)$$

in which b_i is an item location parameter and d_k is a threshold parameter. This decomposition requires that all items have the same number of response categories, which is likely to occur with rating scales, and implies that thresholds are constant across all items.

In the generalized partial credit model (Muraki, 1992) items have different slope parameters, denoted by a

$$P_{ik} = \exp\{\sum_{v=1}^{k} a_i (\theta - b_i + d_v)\} \; / \; \sum_{c=1}^{K} a_i \exp\{\sum_{v=1}^{c} a_i (\theta - b_i + d_v)\}. \qquad (9)$$

The only difference between Equation 7 and Equation 9 is the slope parameter a_i. There is a rating scale version of the generalized model which uses the relation in Equation 8.

Following the approach employed to study item parameter drift in achievement test items (Bock, Muraki, & Pfeiffenberger, 1988), Muraki (1993) proposed a procedure for assessing polytomous DIF using the generalized partial credit model. This DIF procedure posits that the slope parameters are equal in the reference and focal groups, and tests for differences in the step

parameters, b_{ik}. Within the rating scale version of this model, DIF assessment involves checking for differences between the focal and reference group in item location parameter b_i, on an item–by–item basis, and testing across all items for differences in threshold parameters, d_k, across all items. Because the partial credit model is a special case of the generalized partial credit model, in which the item slopes are equal across all items, the same approach can be used to assess DIF for the rating scale model.

Binary Non–Parametric Latent-Variable DIF Procedure

Simultaneous Item Bias Approach (SIBTEST). Currently, this category contains only one latent–variable approach for DIF assessment: the simultaneous item bias test or SIBTEST (Shealy & Stout, 1993a; Shealy & Stout, 1993b). SIBTEST has a theoretical foundation in multidimensional item response theory and bears a close resemblance to the observed–score STND method. A DIF–free multidimensional item response model is postulated to underlie performance on a set of items. The SIBTEST model makes a distinction between the construct of interest or target ability and secondary nuisance abilities, and postulates that DIF for the marginal item response function (IRF) for the target ability results from differences in the distributions of the nuisance abilities between the focal and references groups. In essence, differences along these dimensions introduce construct–irrelevant variance into the measurement process. This model provides a psychometric rationale for the differences in unidimensional item response functions that is consistent with the fundamental distinction between construct relevant and construct irrelevant differences. In the full multidimensional space, each item is DIF–free; differences in distributions of nuisance abilities induce DIF at the unidimensional target ability level.

This latent–variable approach does not posit a particular parametric form for the IRF. Instead, it assesses DIF in the same manner as does STND with one important difference: instead of using the empirical item test regression employed by STND, it regresses item performance onto an estimate based on classical test theory of matching–variable true score.

In SIBTEST, the measure of DIF employed parallels the STD P–DIF index. Differences in the empirical item/true score regressions for the focal and reference groups are averaged across score levels with a focal group weighting function. It has been shown that the true–score correction

improves the matching variable in a way that leads to unbiased estimation of this standardization–like DIF index (Shealy & Stout, 1993b). In SIBTEST, the studied item is not part of the matching variable. In MH and STND the matching variable must include the studied item in order to produce an unbiased estimate (Holland & Thayer, 1988; Donoghue, Holland & Thayer, 1993). A statistical test of the null–DIF hypothesis exists for SIBTEST, as does a standard error for the descriptive index of DIF (Shealy & Stout, 1993b). The SIBTEST approach appears to be as effective for detecting DIF as the MH procedure (Shealy & Stout, 1993a).

Polytomous Non–Parametric Latent-Variable (IRF) DIF Procedure

Polytomous SIBTEST. The latent–variable approach to DIF assessment, called SIBTEST, was actually designed to study differential test functioning (Shealy & Stout, 1993a; Shealy & Stout, 1993b) and is easily adapted to the study of polytomous DIF (Chang, Mazzeo, & Roussos, 1993). Like extended standardization, SIBTEST does not postulate a functional form for the relationship between item scores and scores on the matching variable. Instead of using the empirical item test regression employed by extended STND, it regresses item performance onto an estimate of matching variable true score. Differences in the empirical item/true score regressions for the focal and reference groups are averaged across score levels with a focal group weighting function.

A statistical test of the null–DIF hypothesis exists for SIBTEST, as does a standard error for the descriptive index of DIF (Chang, Mazzeo, & Roussos, 1993; Shealy & Stout, 1993b). In addition, it has been shown that equivalent item response functions (IRFs) or expected item score functions for the focal and reference groups implies equivalent item category response functions for the focal and reference groups, under the partial credit, generalized partial credit, and graded response models (Chang & Mazzeo, 1994). This result suggests that SIBTEST, which assesses DIF with respect to differences in IRFs, can be employed as a first step in testing for polytomous IRT DIF. If DIF is detected, then a latent–variable DIF procedure that posits a particular mathematical form for all item response categories can be used to study the DIF in greater detail. This two–step process mirrors a binary DIF process in which an item is first studied via MH. If an item is flagged for DIF it is then

submitted to STND for distractor analysis in an effort to better understand why the item exhibits DIF (Dorans & Holland, 1993).

In summary, as was the case with the binary DIF procedures, we can cross–classify the polytomous DIF procedures by whether the matching variable is an observable vs. a model–based estimate of an unobservable or latent trait, and whether or not the approach posits a parametric form for the relationship between item score and the matching variable (see Table 2).

Table 2

Cross–Classification for Polytomous DIF Procedures

	A Parametric Form for Relationship Between Item Score and the Matching Variable	**No Parametric Form** for Relationship Between Item Score and the Matching Variable
Observed–Score Matching Variable	Polytomous Logistic Regression	Mantel Polytomous STND HW1 & HW3 Generalized Mantel–Haenszel
Latent–Variable Matching Variable	General IRT–LR Partial Credit Generalized Partial Credit	Polytomous SIBTEST

Evaluation of Polytomous DIF Procedures

Validity of Scoring Rules and Quality of Matching Variable

In the case of binary–scored items, subject matter experts craft an item in such a way that the keyed response is defensible from a content perspective. The item is scored as correct or incorrect, assigned values of 1 and 0; a total score is obtained by summing the item scores (sometimes a correction for guessing is involved) For polytomously–scored items, an obvious extension is to assign arbitrarily consecutive integers to ordered

categories. It is important in the development and administration of this type of item that a sound construct–based reason exist for the assignment of the numerical values to the categories, and that the meaning inferred from the scoring rubric be both reliable and generalizable. When the data are fit by the Rasch model, there is a theoretical justification for the number right score as a matching variable (Holland & Thayer, 1988). Even when the data are not fit by the Rasch model, number correct is still a reasonable matching variable.

The efficacy of DIF assessment also hinges on the quality of the matching variable. In binary–scored DIF analysis, a simple sum of item scores produces a total score that frequently serves as the best available matching variable, i.e., a reliable measure of the construct of interest. There are exceptions to this rule. DIF assessment presumes that all items and the matching variable, be it an observed score or a model-based estimate of ability, are measuring the same dimension. In fact, DIF can be viewed (as it is in the SIBTEST framework) as a violation of unidimensionality. DIF assessment procedures work well as long as violations of unidimensionality are limited. When a test is multidimensional, it may be necessary to decompose the score into more homogeneous subscores, and either match on them separately or use multivariate matching (Dorans & Holland, 1993). Otherwise, the DIF analysis is likely to yield different results in different regions of the multidimensional ability space, as the mix of abilities brought to bear on the item varies. Therefore, when multidimensionality is persavive, DIF is difficult to assess.

For polytomous items, a theoretical justification exists for using number correct if the data follow the partial credit model (Zwick et al., 1993a). For at least one commonly–used polytomous item, the essay, the matching variable issue is complicated by the fact that essays and multiple choice items may measure different dimensions, and that different essays may also measure unique dimensions (Dorans & Schmitt, 1993). When the number of items comprising the matching variable is too small, e.g. less than 20, or if the item being studied is not included in the matching variable, DIF assessment becomes problematic (Donoghue, Holland, & Thayer, 1993). Valid polytomous DIF detection and description requires a well–defined, reliable matching variable.

Statistical and Practical Utility of DIF Procedures

Any DIF procedure can be evaluated in terms of statistical and practical criteria. We propose seven criteria which can be used to assess polytomous DIF procedures. We will, now address each of these criterion in turn, and apply them to a subset of the polytomous DIF procedures that were described in this paper. A cross–classification of the criteria as applied to the selected polytomous DIF procedures is summarized in Table 3.

Statistical Criteria

1. Linkage to test theory. The observed–score approaches, GMH and STND/Mantel, do not make any assumptions about the classical test theory decomposition of scores (Lord & Novick, 1968). Although their binary analogs might be called classical procedures (Scheuneman & Bleistein, 1989), these polytomous observed–score methods have no connection to any theory of tests.

The partial credit model uses the strong true score theory known as IRT, while the polytomous SIBTEST approach uses both a new non–parametric form of IRT and what is in essence traditional test theory (Kelley, 1927). Both these latent–variable approaches are closely linked to a test theory that decomposes an observed score into a systematic true score (or a monotone transformation thereof), and a stochastic error score, both of which are latent variables.

2. Interpretable measure of DIF. To be used effectively, a DIF detection technique needs an interpretable measure of amount of DIF. The definition of DIF varies across polytomous models and thus the complexity of the interpretation of DIF also varies from model to model. Among the methods summarized in the table, the STND/Mantel approach and the polytomous SIBTEST approach have measures of DIF in the metric of expected item score for the focal group, which can be thought of as a weighted difference in empirical item test regressions. For *long* matching variables with high reliability, these approaches should yield identical estimates of amount of DIF. These DIF measures can be interpreted as a difference between how the focal group actually performs on the item and how well matched reference members would have performed on the item.

Table 3

Summary of Selected Polytomous DIF Procedures Evaluated by Suggested Statistical and Practical Criteria

	Generalized Mantel–Haenszel	Polytomous STND	Polytomous SIBTEST	Generalized Partial Credit
Statistical Criteria				
1. Link to test theory	none	none	IRT and CTT bases	IRT basis
2. Interpretable measure of amount of DIF	a set of odds–ratio measures	standardized expected item score measure in focal group metric	standardized expected item score measure in focal group metric	differences in item locations and thresholds in metric of latent variable
3. Unbiasedness of estimator	biased	unbiased	biased	?
4. Standard error	yes	yes	yes	in theory
5. Statistical test of null hypothesis	yes	yes	yes	yes
Practical Criteria				
1. Cost	variable – depends on how many sets of odds ratios are studied	inexpensive	inexpensive	variable – depends on whether tests for DIF are limited to location parameters
2. Capacity to handle multiple items	several items can be analyzed together	several items can be analyzed together	several items can be analyzed together	several items can be analyzed together

The DIF measures that come from the generalized partial credit model are in the metric of the latent ability. These include differences in item locations and differences in thresholds for the rating scale version of the model, and differences in step location for the partial credit version. Unlike the expected item score measures, these measures do not depend directly on the distributions of scores in either the focal or reference group.

The least interpretable measures are associated with the GMH method. The choice of measure depends upon the particular set of comparisons one makes. For example, one possibility is to use the set of odds-ratios which compares each category to a common base category (Zwick, Donoghue, & Grima, 1993a). In many applications, any category could serve as the base, so many possible sets exist. In addition, other types of odds-ratios are possible measures of DIF.

3. Unbiasedness. It is desirable for a DIF procedure to employ a measure of DIF that is unbiased if there is zero DIF. To date, only the polytomous STND measure, as implemented as a supplement to the Mantel statistical test, has exhibited this desirable property, albeit only indirectly via simulation studies (Grima, Zwick, & Donoghue, 1993; Zwick et al., 1993a). The same studies show that the DIF measures for the GMH statistic are biased positively above the null-DIF amount of equal odds. The polytomous SIBTEST measure identifies items favoring the focal group when there is no DIF (Chang, Mazzeo, & Roussos, 1993), which suggests the approach may be overcorrecting for unreliability.

In general the approaches that do not employ a parametric form for the relationship between item scores and the matching variable will tend to be less biased than the approaches that impose a functional form on the data unless the particular functional form is appropriate for the data.

4. Standard error. Ideally, an unbiased estimator also has a low standard error. In practice, however, we often have to choose between an unbiased estimator and one with a smaller standard error. In order to make that choice it is useful to have an estimate of the standard error. Standard errors exist for the both the STND approach and the polytomous SIBTEST approach. The standard error for the GMH procedure has been addressed by (Zwick, Donoghue, & Grima, 1993a). In theory, a standard error exists for the difference in item parameters for the generalized partial credit model. In practice, however, it is computationally demanding because it requires the

inverse of a very large matrix (i.e. dimension equal to the number of items times the number of categories).

In general the approaches that do not employ a parametric form for the relationship between item scores and the matching variable will tend to be less biased than the approaches that impose a functional form on the data unless the particular functional form is appropriate for the data. Biasedness is often a cost associated with imposing a strong model on the data. The benefit of the strong model is the ability to work with weaker data, e.g. data in which there are few items and a short matching variable.

5. Statistical test of null hypothesis. Significance testing of the null DIF hypothesis answers the question of whether the DIF seen on an item exceeds that expected given sampling variability under the null hypothesis. It protects researchers against concluding there is DIF when there is none, but it does not protect them from concluding there is no DIF when there is some. All four procedures provide for a statistical test of the null hypothesis of null–DIF.

The power of a statistical test is the probability that it will lead to the rejection of the null hypothesis, no DIF, in favor of a specific alternative hypothesis. Power increases as a function of sample size and the amount of DIF associated with the effect size (Cohen, 1988). Power is inversely related to the stringency of the significance criteria. The power to detect DIF needs to be studied for the four methods listed in Table 3.

Practical Criteria

1. Cost. The degree to which a DIF procedure will be used in routine operational settings is largely a function of the cost associated with its use. Major cost components include the computer time required, and the human resources needed to check and evaluate the results. When a strong true–score model, such as an IRT model, is used it is important to check whether the model fits the data before making inferences based on the results of using the model in a DIF context. More complex models often require more specialized training to use than do simpler models. In this sense, the GMH is the most costly procedure to use because of the ambiguity associated with definition of its DIF measure. Since the generalized partial credit model may be used to test for differences in thresholds as well as item location parameters, it too can be expensive. In contrast, the STND/Mantel and polytomous SIBTEST

approaches are relatively low cost since each produces a single measure of DIF for each item.

2. Capacity to handle multiple items. Another practical consideration is whether each item must be analyzed separately (as with current implementations of the IRT–LR approach) or whether the items can be analyzed together as a set. A drawback of the flexible General IRT–LR approach is the amount of time required to process each item (Thissen, Steinberg, & Wainer, 1993). Each of the four approaches to polytomous DIF under consideration in this paper (GMH, polytomous STND, polytomous SIBTEST, and generalized partial credit) are capable of studying multiple items simultaneously.

Conclusions

It appears that the polytomous DIF techniques which have been reviewed in this paper are not yet ready for routine operational use. This is not unexpected since these techniques have yet to receive the extensive and rigorous study accorded to their binary DIF counterparts. Among the procedures that we have examined, the expected item score approaches, the polytomous STND(observed score) approach, and the polytomous SIBTEST (latent variable) approach appear closer to practical implementation than either the GMH or generalized partial credit approaches. This is not surprising given that the expected item score approaches collapse across categories via a simple additive rule to arrive at a standard statistical summary of the data, i.e., an expectation. In the process, some valuable information may be lost, but simplicity of interpretation and statistical stability is obtained. In contrast, the GMH approach is more descriptive of the conditional distributions of categorical item scores at each ability level. Descriptions of a set of conditional distributions, however, require more data to achieve a desired level of stability than do estimates of the averages of those distributions. The generalized partial credit model imposes a mathematical structure on these more complex data and gains some stability and interpretability in the process, but is still trying to describe something more complex than an expected item score.

The expected item score approaches, polytomous STND, and polytomous SIBTEST each have the disadvantage of discarding information during DIF analysis because score distributions can not be recreated from averages. However, the focus on averages, or expected item scores, has the advantage of simplicity of interpretation, and provides more statistical stability since averages of conditional distributions are more stable than entire conditional distributions. The GMH approach attempts to summarize these conditional distributions at the expense of interpretational simplicity and statistical stability. The generalized partial credit model imposes a model on the data that reduces the statistical instability, but at the potential expense of employing an inappropriate model for the functional form of the relationship of item score to the latent variable.

For these reasons, it appears unlikely that the GMH and generalized partial credit models are ready for operational use as primary DIF detection devices at the present time. Instead they may be better suited for use as adjuncts to the easier to use and more interpretable polytomous STND and polytomous SIBTEST approaches. The GMH approach is likely to be limited to applications where ample data exists to obtain stable estimates of the conditional item score distributions. The generalized partial credit model, or some other strong true–score model, is more likely to be useful with small samples of data where strong models are needed to extract inferences from the data. Strong true–score models are also more likely to be useful with smaller numbers of items, a situation often encountered in research studies (Bock, 1993).

We propose a framework for classifying DIF procedures on the basis of whether they define DIF with respect to an observed variable or a latent variable, and by whether or not a parametric form is proposed for the relationship between item score and the matching variable. However, the framework does not provide for the inclusion of all potential DIF procedures. Some types of performance assessments (e.g. the portfolio) employ stimuli which produce complex responses. Psychometric models which order examinees and their responses to items are not appropriate for use with these more complex stimuli and resultant responses (Mislevy, 1993). Before DIF procedures can be developed for these kind of stimuli, a "psychometrics for a new generation of tests" must be developed (Mislevy, 1993). The current lack of appropriate psychometric models and DIF procedures to accommodate

these types of stimuli, however, does not exempt these assessment procedures from the need to demonstrate that they are fair and equitable. DIF analysis (or something in this spirit) for performance assessment stimuli which involve scoring schema more complex than ordered polytomous-scoring models, will eventually become a mandated procedure. It seems likely that the solutions to this problem will not be as straightforward as developing polytomous extensions of binary DIF procedures.

Finally, we proposed several criteria that may be used to evaluate the efficacy of DIF procedures. It is our intent that these statistical and practical criteria, in conjunction with our classification framework, will assist practitioners evaluate prospective polytomous DIF procedures in the future.

References

Agresti, A. (1990). *Categorical data analysis.* New York: Wiley.

Allen, N. L., & Holland, P. W. (1993). A model for missing information about the group membership of examinees in DIF studies. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 241–252). Hillsdale, NJ: Lawrence Erlbaum Associates.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43,* 561–573.

Bennett, R. E., & Ward, W. C. (Eds.). (1993). *Construction versus choice in cognitive measurement.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Bock, R. D. (1972). Estimating item parameters and latent ability when the responses are scored in two or more nominal categories. *Psychometrika, 37,* 29–51.

Bock, R. D. (1993). Different DIFs: Comment on the papers read by Neil Dorans and David Thissen. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 115–122). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bock, R. D., & Aikten, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. *Psychometrika, 46,* 442–449.

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika, 35,* 179–197.

Bock, R. D., Muraki, E., & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement, 25,* 275–285.

Camp, R. (1993). The place of portfolios in our changing views of writing assessment. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement* (pp. 183–212). Hillsdale, NJ: Lawrence Erlbaum Associates.

Chang, H-H., & Mazzeo, J. (1994). The unique correspondence of the item response function and the item category response functions in polytomously scored item response models. *Psychometrika, 59.*

Chang, H-H., Mazzeo, J., & Roussos, L. A. (1993, April). *Extension of Shealy–Stout's DIF procedures to polytomously scored items.* Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta, GA.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Donoghue, J. R., Holland, P. W., & Thayer, D. T. (1993). A monte carlo study of factors that affect the Mantel–Haenszel and standardization measures of differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 137–166). Hillsdale, NJ: Lawrence Erlbaum Associates.

Dorans, N. J. (1991, November). *Implications of choice of metric for DIF effect size on decisions about DIF.* Paper presented at the International Symposium on Modern Theories in Measurement: Problems and Issues., Montebello, Quebec, Canada.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel–Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Hillsdale, NJ: Lawrence Erlbaum Associates.

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* (RR–83–9). Princeton, NJ: Educational Testing Service.

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

Dorans, N. J., & Schmitt, A. P. (1993). Constructed response and differential item functioning: A pragmatic perspective. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement* (pp. 135–165). Hillsdale, NJ: Lawrence Erlbaum Associates.

Grima, A. (1993, April). *Extending the Mantel–Haenszel DIF procedure to polytomously scored items.* Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta, GA.

Holland, P. W., & Thayer, D. T. (1985). *An alternative definition of the ETS delta scale of item difficulty* (RR-85-43). Princeton, NJ: Educational Testing Service.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel–Haenszel procedure. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Kelderman, H. (1989). Item bias detection using loglinear IRT. *Psychometrika, 54,* 681–697.

Kelley, T. L. (1927). *The interpretation of educational measurements.* New York: World Book.

Longford, N. T., Holland, P. W., & Thayer, D. T. (1993). Stability of the MH D–DIF statistics across populations. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 171–196). Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison–Wesley.

Madaus, G. F. (1985). Test scores as administrative mechanisms in educational policy. *Phi Delta Kappan, 66,* 611–617.

Mantel, N. (1963). Chi–squares tests with one degree of freedom: Extensions of the Mantel–Haenszel procedure. *Journal of the American Statistical Association, 58,* 690–700.

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

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149–174.

McLaughlin, M. E., & Drasgow, F. (1987). Lord's chi-square test of item bias with estimated and with known ability parameters. *Applied Psychological Measurement, 11,* 161-173.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17,* 297-334.

Miller, T. R., & Spray, J. A. (1993). Logistic discriminant function analysis for DIF identification of polytomously scored items. *Journal of Educational Measurement, 30,* 107-122.

Mislevy, R. J. (1993). Foundations of a new test theory. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 19–39). Hillsdale, NJ: Lawrence Erlbaum Associates.

Morrison, P. (1992). Testing issues in American schools: Issues for research and policy. *Social policy report: Society for Research in Child Development, 6*(2).

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16,* 159–176.

Muraki, E. (1993, April). *Implementing item parameter drift and bias in polytomous item response models.* Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta, GA.

Muraki, E., & Englehard, G. (1989, April). *Examining differential item functioning with BIMAIN.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Muthen, B., & Lehman, J. (1985). Multiple group IRT modeling: Applications to item bias analysis. *Journal of Educational Statistics, 10,* 133–142.

National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform.* Washington, D. C.: U. S. Government Printing Office.

National Educational Goals Panel. (1991). *The national educational goals report: Building a nation of learners.* Washington, D. C.: Author.

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve problems. *Psychometrika, 56,* 611-630.

Rogers, H. J., & Swaminathan, H. (1993, April). *Differential item functioning procedures for non–dichotomous responses.* Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta, GA.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph No. 17, 34* (4 Part 2). Richmond, VA: William Byrd Press.

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

Scheuneman, J. D., & Bleistein, C. A. (1989). A consumer's guide to statistics for identifying differential item functioning. *Applied Measurement in Education*, 2, 255–275.

Shealy, R. T., & Stout, W. F. (1993a). An item response model for test bias and differential test functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 197–239). Hillsdale, NJ: Lawrence Erlbaum Associates.

Shealy, R. T. , & Stout, W. F., (1993b). A model–based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika, 54*, 159-194.

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

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, 27, 361–370.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 147–169). Hillsdale, NJ: Lawrence Erlbaum Associates.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–113). Hillsdale, NJ: Lawrence Erlbaum Associates.

Thissen, D., & Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika, 49*, 501–519.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51*, 567–577.

Thissen, D., & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika, 56*, 611-630.

U.S. Congress, Office of Technology Assessment. (1992). *Testing in American schools: Asking the right questions* (OTA–SET–519). Washington, DC: U.S. Government Printing Office.

Valencia, S. W., & Calfee, R. (1991). The development and use of literacy portfolios for students, classes, and teachers. *Applied Measurement in Education, 4*, 333–345.

Wainer, H. (1993). Model–based standardized measurement of an item's differential impact. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 123–135). Hillsdale, NJ: Lawrence Erlbaum Associates.

Wainer, H., Sireci, S. G., & Thissen, D. (1991). Differential testlet functioning: Definitions and detection. *Journal of Educational Measurement, 28*, 197–219.

Welch, C., & Hoover, H. D. (1993). Procedures for extending item bias techniques to polytomously scored items. *Applied Measurement in Education , 6*, 1-19.

Wilson, A. W., Spray, J. A., & Miller, T. R. (1993, April). *Logistic regression and its use in detecting nonuniform differential item functioning.* Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta, GA.

Zwick, R., Donoghue, J. R., & Grima, A. (1993a). *Assessing differential item functioning in performance tasks.* (RR-93-14). Princeton, NJ: Educational Testing Service.

Zwick, R., Donoghue, J. R., & Grima, A. (1993b). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30*, 233-251.