DIF Assessment for Polytomously Scored Items: A Framework for Classification and Evaluation

Maria T. Potenza and Neil J. Dorans

Educational Testing Service

Increased use of alternatives to the traditional dichotomously scored multiple-choice item yield complex responses that require complex scoring rules. Some of these new item types can be polytomously scored. DIF methodology is well-defined for traditional dichotomously scored multiple-choice items. This paper provides a classification scheme of DIF procedures for dichotomously 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 dichotomous DIF technique is presented. Several polytomous DIF techniques are evaluated in terms of statistical and practical criteria. *Index terms: DIF methodology, differential item functioning, item bias, polytomous scoring, statistical criteria for differential item functioning.*

Differential item functioning (DIF) refers to a psychometric difference in the way a test 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 trait levels.

The vast majority of multiple-choice tests score items as correct or incorrect. Even when a multiple-choice test is not scored this way, it is often analyzed as if it were (Dorans, 1991). Most procedures used to assess DIF presume that items are scored in this dichotomous fashion (Holland & Wainer, 1993). Currently there are numerous methods for conducting DIF assessment for dichotomously scored items (see Millsap & Everson, 1993, and Scheuneman & Bleistein, 1989, for reviews).

Educational reform efforts have led to an increased use of alternatives to the traditional dichotomously scored multiple-choice item. Many of the stimuli used 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, 1995; Grima, 1993; Muraki, 1993; Rogers & Swaminathan, 1993; Welch & Hoover, 1993; Wilson, Spray, & Miller, 1993; Zwick, Donoghue, & Grima, 1993b). However, several important methodological issues must be addressed in the transition from dichotomous 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.

This paper has three goals. First, a classification scheme for DIF procedures used with dichotomously scored items is suggested, and this classification system is applied to DIF procedures for polytomously scored items. Second, several issues are delineated associated with the extension of current DIF procedures to performance assessments in which polytomous scoring rules are used. Finally, criteria are proposed for the evalua-

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 19, No. 1, March 1995, pp. 23–37

0146-6216/95/010023-15$2.00

tion of polytomous DIF techniques and a selected set of polytomous DIF techniques are evaluated in terms of these criteria.

A Framework for the Classification of DIF Procedures

Two classes of DIF procedures exist for dichotomous 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 observed score as the matching variable, but the latter uses an estimate of latent trait level, 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, procedures are distinguished that use 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 that define a functional form for the relationship between item score and matching variable, and those that do not. Consequently, the false impression may be conveyed that all latent variable models use a parametric form, and that all observed score approaches do not. The framework presented here adds this important distinction and can be used to classify both dichotomous 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 DIF that is detected may be an artifact of model misspecification. In addition, very large sampling covariation among parameter estimates is often a problem for parametric approaches that use several parameters (Lord, 1980; Ramsay, 1991; Thissen & Wainer, 1982). Although the nonparametric procedures are relatively free of model misspecification and colinearity problems, they require sufficient data to directly estimate the regressions of item score on test score. In small samples, these procedures may produce unstable results due to the effects of sampling error.

These two definitional distinctions—matching on an observable versus 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. Table 1 shows this cross-classification with methods that have been proposed for detecting DIF in both dichotomous and polytomous items.

Observed Score DIF Procedures

Although there are many observed score procedures for assessing DIF on dichotomously scored items (Holland & Wainer, 1993; Scheuneman & Bleistein, 1989), three methods are described here for which initial attempts have been made to extend these procedures to the case of polytomous DIF. These are 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 use an observed score measure of the construct of interest as a matching variable. Hence, they state that there is no DIF between groups after they have been matched on an observed score, which is usually the total score. None of these three ob-

Table 1
Cross-Classification of Dichotomous and Polytomous DIF Procedures According to Whether They are Parametric (A Parametric Form is Assumed for the Relationship Between Item Score and the Matching Variable) or Nonparametric (No Parametric Form is Asssumed for the Relationship Between Item Score and the Matching Variable)

Type of Procedure and Matching Variable	Parametric	Nonparametric
Dichotomous DIF		
Observed Score	LRDIF	MH STND
Latent Variable	General IRTLR Limited Information IRTLR Loglinear IRTLR IRT-D^2 Lord's χ^2	SIBTEST
Polytomous DIF		
Observed Score	Polytomous LRDIF	$MNTL_{P\text{-}DIF}$ Polytomous STND HW1 HW3 GMH
Latent Variable	General IRTLR PCM	Polytomous SIBTEST GPCM

served score methods postulates 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).

Dichotomous Nonparametric Observed Score DIF Procedures

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

Standardization. The null DIF definition for the STND method states that at each level of the matching variable there is no difference in proportion correct between the focal group (the focus of the DIF analysis, e.g., the minority group) and the reference group (the basis for comparison, e.g., the majority group). This can be expressed 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, $STND_{P\text{-}DIF}$ [referred to as STD P-DIF by Dorans & Holland (1993)], 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 (SE) 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 STND approach, although a test statistic involving the ratio of $STND_{P\text{-}DIF}$ to its SE can be used.

Mantel-Haenszel. The null DIF definition for the MH method is that the odds for responding correctly

are 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 the definition of null DIF for the STND approach, which is in terms of proportion 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 the amount of DIF under the restriction that the odds-ratio is 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 & Haenszel (1959) provided an estimate of the constant odds-ratio (α_{MH}) that ranges from 0 to ∞: A value of 1 indicates null DIF. In general, odds are converted to log odds because the latter is symmetric around 0 and easier to interpret. Holland & Thayer (1985) converted α_{MH} into a difference in Δ, called $MH_{D\text{-}DIF}$, using a log-odds transformation. The SE of $MH_{D\text{-}DIF}$ has been studied extensively, and $MH_{D\text{-}DIF}$ performs well in detecting DIF (Donoghue et al., 1993). In addition, the MH procedure uses a χ^2 significance test (Mantel & Haenszel).

Polytomous Nonparametric Observed Score DIF Procedures

STND and MH are closely related dichotomous DIF techniques that measure DIF identically when generalized to the polytomous situation. One generalization of the MH procedure is the Mantel (1963) procedure, which was developed to detect association between matched groups on ordinal variables. 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). The STND model will be extended to polytomously scored items below, and its relationship to the polytomous DIF version of the Mantel procedure will be discussed.

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 dichotomous and polytomous DIF is the number of levels of the dependent variable (i.e., the item score). For dichotomous items, the $STND_{P\text{-}DIF}$ index is an average weighted difference in proportion correct (the expected item score under dichotomous scoring) across score levels. The more general index is $STND_{ES\text{-}DIF}$, or standardized expected item score DIF. For the general case, assume that there is: (1) a matching variable, X, with M levels, $m = 1, \ldots, M$; (2) an ordered item score, Y, with K levels or categories, $k = 1, \ldots, K$; and (3) two groups—the reference (R) and the focal (F) groups.

For the polytomous version of STND, first compute expected item scores for both groups—$E_{Fm}(Y|X)$ and $E_{Rm}(Y|X)$—using

$$E_{Fm}(Y|X) = \sum_k N_{Fmk}\, Y_k / N_{Fm} \tag{1}$$

and

$$E_{Rm}(Y|X) = \sum_k N_{Rmk}\, Y_k / N_{Rm}\,, \tag{2}$$

where

N_{Fmk} is the number of examinees in the focal group at score level m with item score Y_k,
N_{Fm} is the total number of examinees in the focal group at score level m,
N_{Rmk} is the number of examinees in the reference group at score level m with item score Y_k, and

N_{Rm} is the total number of examinees in the reference group at score level m.
The item score variable, Y_k, can take on any ordered values, including 1, 2, 3, ..., K.

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

$$D_m = E_{Fm}(Y|X) - E_{Rm}(Y|X), \tag{3}$$

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

$$STND_{ES\text{-}DIF} = \sum_m N_{Fm} D_m / N_F, \tag{4}$$

where N_F is the total number of focal group examinees.

HW1 and HW3 approaches. Recently, another pair of test statistics, HW1 and HW3, have been proposed for detecting departures from null DIF for polytomously scored items (Welch & Hoover, 1993). Both indexes can be described using the STND framework.

For HW1, the difference in expected item performance at each level of the matching variable is computed and then converted into a t statistic by dividing by a pooled SE 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 the independent t statistics. The resultant statistic is normally distributed with a mean of 0 and a standard deviation of 1.

HW3 weights each test statistic by the reciprocal of its sampling variance. A correction factor is used 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 depend on sample size so they are not measures of the amount of DIF.

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

$$MNTL_{P\text{-}DIF} = \left[\sum_m F_m - \sum_m E(F_m)\right]^2 \Big/ \sum_m Var(F_m), \tag{5}$$

where

$$F_m = E_{Fm}(Y|X) N_{Fm}, \tag{6}$$

and $E(F_m)$ and $Var(F_m)$ are the mean and variance, respectively, of F_m under the null hypothesis of no association between group and item score given the value of the matching variable. Under the null hypothesis, $MNTL_{P\text{-}DIF}$ is distributed as a χ^2 with 1 degree of freedom (Mantel, 1963; Zwick et al., 1993b).

Generalized MH approach. The generalized MH (GMH) procedure is another generalization of the dichotomous MH procedure (Mantel & Haenszel, 1959). Whereas the polytomous STND and Mantel procedure emphasize expected (average) item scores when comparing focal and reference groups, the 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 define 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 reference groups; 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.

Dichotomous Parametric Observed Score DIF Procedure

Logistic regression approach. The LRDIF approach is an observed score method that specifies a particular parametric form for the relationship between item score and matching variable. Swaminathan & Rogers (1990) postulated 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 postulated a parametric functional form for the empirical regression used by STND. Significance tests exist for both uniform DIF and nonuniform DIF (i.e., item-test regressions with intersection points).

The MH procedure also 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 0 (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 DIF 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 of the magnitude of DIF, $MH_{D\text{-}DIF}$ and $STND_{P\text{-}DIF}$, respectively, that are well studied (Allen & Holland, 1993; Donoghue, Holland, & Thayer, 1993; Dorans & Holland, 1993; Longford, Holland, & Thayer, 1993). Swaminathan & Rogers (1990) did not propose a descriptive statistic for the 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 applied in many ways. 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 different sets of odds ratios.

Latent Variable DIF Procedures

A second class of DIF techniques for dichotomous items is based on either 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 the latent trait, the underlying proficiency, or the true score. A fundamental difference between the latent variable approaches and the observed score approaches is the use of estimates, derived from observed data, of the latent trait 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 (IRF).

Dichotomous Parametric Latent Variable DIF Procedures

Parametric IRT approaches 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 IRF assumed, the type of parameter estimation used, 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 (IRTLR; Thissen, Steinberg, & Wainer, 1988), which uses the Bock-Aitkin (Bock & Aitkin, 1981) marginal maximum likelihood estimation algorithm to estimate parameters for a number of different IRT models. A second approach, loglinear IRTLR, uses maximum likelihood estimation (Kelderman, 1989). A third approach, limited-information IRTLR, uses normal ogive IRT models with generalized least squares estimation of parameters (Muthén & Lehman, 1985). Each of these three approaches use 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 LR approaches have been evaluated by Thissen et al. (1993). The least applicable for DIF analysis for traditional dichotomously scored multiple-choice items is the loglinear IRTLR procedure, because it is restricted to Rasch models which do not permit items to have different discriminations or nonzero lower asymptote parameters. The normal ogive IRT models used by the limited-information IRTLR approach similarly do not permit nonzero lower asymptotes and require larger sample sizes than the other LR methods, but they can be used to test for DIF within a multidimensional model. Because the general IRTLR 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 IRTLR approach, because each item is studied separately and because 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, IRT-D^2, analyzes all of the items simultaneously. IRT-D^2 uses the Bock-Aitkin marginal maximum likelihood EM algorithm followed by one or two iterations of the Bock & Lieberman (1970) direct Newton-Raphson algorithm to estimate item parameters for the reference and focal groups (Bock, Muraki, & Pfeiffenberger, 1988). The Newton-Raphson algorithm provides SEs for the item parameter estimates. The three-parameter logistic model is 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 SEs to evaluate the significance of observed differences.

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

Lord (1980) also suggested a procedure within this category, which is known as Lord's χ^2 approach (McLaughlin & Drasgow, 1987). As with the IRT-D^2 method, Lord's χ^2 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 χ^2 test is used to simultaneously test the null hypothesis of no differences in both parameters across groups.

Polytomous Parametric Latent Variable 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 selecting each category as a function of underlying

trait level. 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 selecting 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 example of a difference model (Samejima, 1969). For the divide-by-total class of models, the parametric form for the probability of selecting 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 (PCM; Masters, 1982), and the rating scale model (Andrich, 1978) are divide-by-total models. The multiple-choice model is a modified version of the nominal model that allows for nonzero 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 used to study polytomous DIF. The approach uses a series of 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 PCM (Masters, 1982), the propensity to select category k for item i is expressed as

$$P_{ik} = \exp\left[\sum_{v=1}^{k}(\theta - b_{iv})\right] \Big/ \sum_{c=1}^{K}\exp\left[\sum_{v=1}^{c}(\theta - b_{iv})\right], \tag{7}$$

where the b_{iv} are the points of intersection for adjacent category response functions, called step parameters, and θ is the individual's trait level.

In the PCM, all items have the same discrimination parameter (set equal to 1.0); therefore, it does not appear in Equation 7. This makes the PCM an extension of the dichotomous Rasch model.

The rating scale model (Andrich, 1978) can be derived from the PCM by decomposing

$$b_{ik} = b_i - d_k\,, \tag{8}$$

where 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 PCM (GPCM; Muraki, 1992), items can have different slope parameters, denoted by

$$P_{ik} = \exp\left[\sum_{v=1}^{k} a_i(\theta - b_i + d_v)\right] \Big/ \sum_{c=1}^{K}\exp\left[\sum_{v=1}^{c} a_i(\theta - b_i + d_v)\right]. \tag{9}$$

The only difference between Equation 7 and Equation 9 is that Equation 9 includes the slope parameter a_i. There is a rating scale version of the generalized PCM that uses the relation in Equation 8 (Andrich, 1978).

Following the approach used to study item parameter drift in achievement test items (Bock et al., 1988), Muraki (1993) proposed a procedure for assessing polytomous DIF using the GPCM. 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}. For the rating scale version of this model, DIF is assessed by evaluating differences between the focal and reference group on the item location parameter, b_i, on an item-by-item basis, and testing across all items for differences in threshold parameters, d_k. Because the PCM is a special case of the GPCM, in which the item slopes are equal across all items, the same approach can be used to assess DIF for the PCM.

Dichotomous and Polytomous Nonparametric Latent Variable DIF Procedures

The simultaneous item bias test (SIBTEST; Shealy & Stout, 1993a, 1993b) has a theoretical foundation in

multidimensional IRT 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. SIBTEST distinguishes between the construct of interest or target trait and secondary nuisance traits, and postulates that DIF for the marginal IRF for the target trait results from differences in the distributions of the nuisance traits between the focal and reference 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 IRFs 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 traits induce DIF at the unidimensional target trait level.

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

In SIBTEST, the measure of DIF used parallels the $STND_{P\text{-}DIF}$ 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. The true score correction improves the matching variable in a way that leads to unbiased estimation of this STND-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 (Donoghue et al., 1993; Holland & Thayer, 1988). A statistical test of the null DIF hypothesis exists for SIBTEST, as does a SE 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).

SIBTEST actually was designed to study differential test functioning (Shealy & Stout, 1993a, 1993b) and is easily adapted to the study of polytomous DIF (Chang et al., 1995). Equivalent IRFs or expected item score functions for the focal and reference groups imply equivalent item category response functions for the focal and reference groups under the PCM, GPCM, and graded response models (Chang & Mazzeo, 1994). This suggests that SIBTEST, which assesses DIF with respect to differences in IRFs, can be used 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 is similar to a dichotomous DIF process in which an item is first studied using MH. If an item is identified as a DIF item, it is then analyzed by the STND method for distractor analysis in an effort to better understand why the item exhibits DIF (Dorans & Holland, 1993).

Evaluation of Polytomous DIF Procedures

Validity of Scoring Rules and Quality of Matching Variable

For dichotomously scored items, content experts construct an item so that the keyed response is defensible from a content perspective. The item is scored as correct or incorrect and assigned values of 1 and 0, respectively. 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 method be both reliable and generalizable. When the data fit the Rasch model, there is a theoretical justification for the number-correct score as a matching variable (Holland & Thayer, 1988). Even when the data do not fit the Rasch model, number correct may be a reasonable matching variable.

The efficacy of DIF assessment also hinges on the quality of the matching variable. In dichotomously 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 (an observed score or a model-based estimate of trait level) measure 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 trait space, as the mix of traits brought to bear on the item varies. Therefore, when multidimensionality is pervasive, DIF is difficult to assess.

For polytomous items, a theoretical justification exists for using number correct if the data follow the PCM (Zwick, Donoghue, & Grima, 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 et al., 1993). Valid polytomous DIF detection and description requires a well-defined, reliable matching variable.

Statistical and Practical Utility of DIF Procedures

Seven criteria are proposed that can be used to assess polytomous DIF procedures. Each of these criteria is discussed and applied to a subset of the polytomous DIF procedures. Criteria 1–5 are statistical criteria, and criteria 6–7 are practical criteria. A cross-classification of the criteria as applied to selected polytomous DIF procedures is summarized in Table 2.

Table 2
Summary of Selected Polytomous DIF Procedures Evaluated by Suggested Statistical and Practical Criteria

Criteria	GMH	Polytomous STND	Polytomous SIBTEST	GPCM
Statistical Criteria				
1. Link to test theory	none	none	IRT and CTT	IRT
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	unknown
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

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 dichotomous analogues might be called classical procedures (Scheuneman & Bleistein, 1989), these polytomous observed score methods have no connection to any test theory.

   The PCM uses IRT strong true score theory; the polytomous SIBTEST approach uses both a new nonparametric form of IRT and what is in essence classical test theory (Kelley, 1927). Both of 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 the 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 Table 2, 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 matching variables based on long tests with high reliability, these approaches should yield identical estimates of the 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 group members would have performed on the item.

   The GPCM DIF measures are in the latent trait metric. 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 PCM 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 on the particular set of comparisons made. For example, one possibility is to use the set of odds ratios that compare each category to a common base category (Zwick et al., 1993a). In many applications, any category could serve as the base; therefore, 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 use a measure of DIF that is unbiased if there is no DIF. Only the polytomous STND measure, as implemented as a supplement to the Mantel statistical test, has exhibited this desirable property using simulation studies (Grima, 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 et al., 1995), which suggests that the approach may be overcorrecting for unreliability.

   In general, the approaches that do not use a parametric form for the relationship between item scores and the matching variable 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 matching variable based on a short test).
4. *Standard error.* Ideally, an unbiased estimator also has a small SE. In practice, however, a choice must be made between an unbiased estimator and one with a smaller SE. In order to make that choice it is useful to have an estimate of the SE. SEs exist for both the STND approach and the polytomous SIBTEST approach. The SE for the GMH procedure has been addressed by Zwick et al. (1993a). In theory, a SE exists for the difference in item parameters for the GPCM. 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 procedures that do not impose a functional form on the relationship of item scores to

the matching variable tend to exhibit more variability in their measures of DIF. One of the advantages of imposing a functional form is a reduction in variability, which often comes at the expense of increased statistical bias in the estimator.

5. *Statistical test of the null hypothesis.* Significance testing of the null DIF hypothesis answers the question of whether the DIF exceeds that expected given sampling variability under the null hypothesis. It protects researchers against concluding that there is DIF when there is none, but it does not protect them from concluding that 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 of 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 2.
6. *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 evaluate the results. When a strong true-score model, such as an IRT model, is used it is important to evaluate 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 the definition of its DIF measure. Because the GPCM 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 because each produces a single measure of DIF for each item.
7. *Capacity to handle multiple items.* Another practical consideration is whether each item must be analyzed separately (as with current implementations of the IRTLR approach) or whether the items can be analyzed together as a set. A drawback of the flexible general IRTLR approach is the amount of time required to process each item (Thissen et al., 1993). Each of the four approaches to polytomous DIF listed in Table 2 is capable of studying multiple items simultaneously.

Conclusions

The polytomous DIF techniques that have been reviewed here are not yet ready for routine operational use. This is not unexpected because these techniques have yet to receive the extensive and rigorous study accorded to their dichotomous DIF counterparts. The expected item score approaches, the polytomous STND (observed score) approach, and the polytomous SIBTEST (latent variable) approach appear to be closer to practical implementation than either the GMH or GPCM. This is not surprising given that the expected item score approaches collapse across categories using 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 trait 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 GPCM imposes a mathematical structure on these more complex data and gains some stability and interpretability in the process, but it still attempts 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 cannot 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 because 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 GPCM imposes a model on the data that reduces statistical instability, but at the potential expense of using 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 GPCM are ready for operational use as primary DIF detection procedures 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 in which ample data exist to obtain stable estimates of the conditional item score distributions. The GPCM, or some other strong true-score model, is more likely to be useful with small samples of data in which 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).

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 describes the relationship between item score and the matching variable, was proposed. However, the framework does not provide for the inclusion of all potential DIF procedures. Some types of performance assessments (e.g., the portfolio) use stimuli that produce complex responses. Psychometric models that 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, new psychometric models (e.g., Mislevy, 1993) must be developed. 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 that involve scoring schema more complex than ordered polytomous scoring models will eventually become necessary. The solutions to this problem likely will not be as straightforward as developing polytomous extensions of dichotomous DIF procedures.

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: Erlbaum.

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

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: Erlbaum.

Bock, R. D., & Aitkin, 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.

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,* 391–404.

Chang, H., Mazzeo, J., & Roussos, L. A. (1995). *Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure* (Research Rep. No. 95-5). Princeton NJ: Educational Testing Service.

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

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: Erlbaum.

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: Erlbaum.

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: Erlbaum.

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: Erlbaum.

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

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: Erlbaum.

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

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

Mantel, N. (1963). Chi-square 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.

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.

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

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: Erlbaum.

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.

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

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.

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 Edu-*

cation, 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: Erlbaum.

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.

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. (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., 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: Erlbaum.

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: Erlbaum.

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

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: Erlbaum.

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.

Acknowledgments

The authors 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 earlier versions 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.

Author's Address

Send requests for reprints or further information to Maria T. Potenza, Mail Stop 32-V, Educational Testing Service, Princeton NJ 08541, U.S.A. Internet: mpotenza@ets.org.