ED 390 886 TM 024 167

AUTHOR	Chang, Hua-Hua; And Others
TITLE	Detecting DIF for Polytomously Scored Items: An Adaptation of the SIBTEST Procedure. Research Report.
INSTITUTION	Educational Testing Service, Princeton, N.J.
SPONS AGENCY	Office of Educational Research and Improvement (ED), Washington, DC.
REPORT NO	ETS-RR-95-5
PUB DATE	Jan 95
CONTRACT	R999G30002
NOTE	35p.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Evaluation/Methods; *Identification; *Item Bias; *Robustness (Statistics); Simulation; *Test Items
IDENTIFIERS	Dichotomous Scoring; Mantel Haenszel Procedure; Partial Credit Model; Polytomous Scoring; *SIBTEST (Computer Program); Standardized Mean Difference; Type I Errors

ABSTRACT

Recently, R. Shealy and W. Stout (1993) proposed a procedure for detecting differential item functioning (DIF) called SIBTEST. Current versions of SIBTEST can only be used for dichotomously scored items, but this paper presents an extension to handle polytomous items. The paper presents: (1) a discussion of an appropriate definition of DIF for polytomously scored items; (2) a modified SIBTEST procedure for detecting DIF for polytomously scored items; and (3) the results of two simulation studies comparing the modified SIBTEST with the Mantel-Haenszel and standardized mean difference procedures. One study used data constrained by the Rasch-like partial credit model, and the other used data with distinct discriminations across items. The simulations indicate that the methodology of including the studied item in matching subtests for controlling impact-induced Type I error tends to yield Type I and Type II error inflation rates that are highly unacceptable when the equal discrimination condition is violated. They provide evidence that the modified SIBTEST procedure is more robust with regard to controlling impact-induced Type I error rate inflation than the other procedures. (Contains 1 figure, 6 tables, and 26 references.) (Author/SLD)

RESEARCH REPORT

DETECTING DIF FOR POLYTOMOUSLY SCORED ITEMS: AN ADAPTATION OF THE SIBTEST PROCEDURE

Hua-Hua Chang
John Mazzeo
Louis Roussos

Educational Testing Service
Princeton, New Jersey
January 1995

Detecting DIF for Polytomously Scored Items: An Adaptation of the SIBTEST Procedure

Hua-Hua Chang and John Mazzeo
Educational Testing Service
Louis Roussos
University of Illinois at Urbana-Champaign

December 8, 1994

This research was partially supported by (a) Educational Testing Service Allocation Projects No.79409, 79413; and (b) the National Assessment of Educational Progress (Grant No. R999G30002) as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

The authors wish to thank Rebecca Zwick, Dorothy Thayer, Eugene Johnson, Howard Wainer, Charlie Lewis, Kentoro Yamamoto, Spence Swinton, Hsin-hung Li, and especially John Donoghue and William Stout for their useful discussions and assistance. The authors wish to particularly thank Nancy Petersen and Ming-Mei Wang for enlightening suggestions in their through review. Comments or requests for reprints should be sent to either of the three authors.

ABSTRACT

Recently, Shealy and Stout (1993) proposed a DIF detecting procedure SIBTEST, which is 1) IRT model based, 2) non-parametric, 3) does not require IRF estimation, 4) provides a test of significance, and 5) estimates the amount of DIF. Current versions of SIBTEST can only be used for dichotomously scored items. However, in this paper an extension to handle polytomous items is developed. This paper presents: (1) a discussion of an appropriate definition of DIF for polytomously scored items, (2) a modified SIBTEST procedure for detecting DIF for polytomous items, and (3) the results of two simulation studies comparing the modified SIBTEST with the Mantel and SMD procedures, one study with data constrained by the Rasch-like partial credit model (same discrimination across polytomous items), and the other study with data having distinctly discrimations across items. These simulation studies indicate that the methodology of *including the studied item in matching subtest* for controling impact induced (group ability differences existing) Type I error tends to yield Type-I/Type II error inflation rates that are highly unacceptable when the *equal discrimination* condition is violated. These simulation studies provide compelling evidence that the modified SIBTEST procedure is much more robust with regard to controlling impact-induced Type I error rate inflation than the other procedures.

Key words: item bias, differential item functioning, DIF, item response theory, polytomous item, partial credit model, generalized partial credit model, graded response model, invariance, ordered categories, SIBTEST.

1. Introduction

The increased use of ordinally scored polytomous items in educational achievement tests has generated considerable interest in the development of DIF detection methods for such items (Dorans & Schmitt, 1993; Miller & Spray, 1993; Welch & Hoover, 1993; Zwick, Donoghue, & Grima, 1993).

For binary scored items a variety of approaches for detecting differential item functioning (DIF) have been well established (see the volume edited by Holland & Wainer, 1993, for an up-to-date review of a number of these approaches). Dorans and Potenza (1994) recently proposed a two-dimensional framework for classifying these procedures. On one dimension, they distinguish between procedures that use an *observed score* as a matching variable versus procedures that match groups in terms of an estimate of a *latent variable.* On the second dimension, they distinguish between *parametric* approaches that assume a parametric functional form for the item response function versus procedures that do not make such assumptions, i.e., *non-parametric* approaches. Examples of *observed-score/non-parametric* approaches are the standardization method (Dorans & Kulick, 1986) and the Mantel-Haenszel procedure (Holland & Thayer, 1988). An example of an *observed-score/parametric* approach is a logistic regression procedure (Rogers & Swaminathan, 1990). Examples of *latent-variable/parametric* procedures are the item response theory (IRT) methods discussed in Thissen, Steinberg, and Wainer (1993). The SIBTEST procedure proposed by Shealy and Stout (1993) was described as a *latent-variable/non-parametric* approach. To avoid confusion, we note here that the observed score is used by SIBTEST as part of the process of achieving groups matched on a latent variable.

Observed-score/non-parametric generalizations for ordinally scored polytomous items have already been suggested by Dorans and Schmitt (1993), Welch and Hoover (1993), and Zwick et.al. (1993). The existence and use of item response models like the partial credit model (Masters, 1982), the generalized partial credit model (Muraki, 1992), and the graded-response model (Samejima, 1969) offer the opportunity for generalizations of *latent-variable/parametric* approaches.

The current paper describes a relatively simple extension of the *latent-variable/non-parametric* SIBTEST procedure for use with polytomously scored items. The paper is organized as follows. First, a discussion of an appropriate definition of DIF for polytomously scored items is presented. Next, the modified SIBTEST procedure is described. An ensuin. .ection presents two simulation studies which evaluate the performance of SIBTEST by comparing its performance to two of the *observed-score/non-parametric* ordinal polytomous procedures reported on to date in the literature. The first study compares the modified SIBTEST procedure with the Mantel procedure (Zwick et al., 1993) and the standardized mean difference (SMD) procedure (Dorans & Schmitt, 1993) using data constrained by the partial credit model (same discrimination across polytomous items). The second study compares the three procedures for use with data having distinct discrimination across items. The final section discusses directions for future work.

2. Conceptualization of DIF for Ordinal Polytomous Items

A Review of Dichotomous DIF Modeling

Observed-score ΓIF methods for binary scored items assume a common definition of null DIF that an item does not exhibit DIF if the regression of item score on matching score is identical for the groups under study.

Definition 1 *(Observed Score Version of Null-DIF) Let Y be the score of an item under study and X be the matching score. Denote $E_R[Y|X]$ and $E_F[Y|X]$ as regressions of Y on X for reference and focal groups, respectively. An item does not exhibit DIF if*

$$E_R[Y|X] = E_F[Y|X] \quad \textit{for all values of } X.$$

Latent-variable methods (both non-parametric and parametric) assume that an item does not exhibit DIF if its latent variable IRFs are identical for the groups under study.

Definition 2 *(Latent Variable Version of Null-DIF) Let $E_R[Y|\theta]$ and $E_F[Y|\theta]$ denote the regressions of Y on latent variable θ for reference and focal groups, respectively. An* *item does not exhibit DIF if*

$$E_R[Y|\theta] = E_F[Y|\theta] \quad \textit{for all values of } \theta. \tag{1}$$

(Note that in the IRT literature $E_R[Y|\theta]$ and $E_F[Y|\theta]$ are referred to as item response functions or IRFs.) The observed-score and latent-variable null DIF definitions are not in general equivalent, and several authors (Holland & Thayer, 1988; Zwick, 1990; and, Meredith & Millsap, 1992) have discussed specific conditions required for their equivalence.

A Definition of DIF for Polytomously Scored Items

Observed-score/non-parametric procedures for ordinal items, such as Dorans and Schmitt's (1993) standardization based approach and the Mantel procedure suggested by Zwick et. al. (1993), adopt a definition of null DIF that is analogous to Definition 1. In these approaches an item exhibits null DIF if the regression of ordinal item score on the matching test score is identical for the groups under study. One possible generalization of an IRT latent variable definition of null DIF for ordinal polytomous items would require that the regression of ordinal item score on the latent variable be identical for the groups under study.

Let Y be the score on the studied item. Assume Y can be scored in terms of $m+1$ ordered categories (e.g., $Y = k$, $0 \leq k \leq m$). Let $P_{k,g}(\theta)$ denote the item category response function (ICRF), the probability of getting score k for a randomly sampled examinee with proficiency θ from group g ($g = R$ or F). The regression of item score on ability can be defined as a weighted sum of ICRFs:

$$E_g[Y|\theta] = \sum_{k=1}^{m} kP_{k,g}(\theta). \tag{2}$$

In analogy to the binary case, we will refer to this regression as an item response function (IRF). It should be noted that binary scored items can also be viewed as ordinal polytomous items with $m = 1$. In such cases $P_{1,g}(\theta) = E_g[Y|\theta]$ and $P_{0,g}(\theta) = 1 - P_{1,g}(\theta)$.

For dichotomous IRT models the structure of an item is obviously completely determined by the specification of its IRF. In other words, $E_g[Y|\theta]$ corresponds to a unique set of ICRFs. Equality of IRFs implies equality of ICRFs. However, for polytomous models, the item structure is in general only completely determined if all m ICRFs are specified. More specifically, it is not guaranteed that a unique correspondence exists between an IRF and a set of ICRFs. If equation (1) does not in general imply

$$P_{kR}(\theta) = P_{kF}(\theta), \quad k = 1, \ldots, m, \quad \textit{for all } \theta, \tag{3}$$

where $P_{kR}(\theta)$ and $P_{kF}(\theta)$ are ICRFs for reference and focal groups respectively, one might prefer to use (3) as the latent variable Null-DIF definition. Chang and Mazzeo (1994) have proved that Equation (1) does indeed imply Equation (3) for three of the most commonly used ordinal item response models – the partial credit model (Masters, 1982), the generalized partial credit model (Muraki, 1992), and the graded response model (Samejima, 1969). Therefore, for the most commonly used polytomous parametric IRT models, a unique correspondence does exists indeed between an IRF and a set of ICRFs. Assessing DIF only with respect to differences in IRFs across ability groups suffices in such cases; no information is lost by comparing only the IRFs. In what follows, Definition 2 is adopted as our latent variable IRF definition of null DIF for ordinally scored polytomous items.

3. An Extension of SIBTEST for Polytomous Items

Shealy-Stout's Multidimensional Viewpoint

SIBTEST, like many other procedures, defines non-DIF in terms of invariant target ability IRFs across groups. Heuristically, Shealy and Stout (1993) model DIF (which for technical reasons they sometimes call "bias") by postulating a multidimensional latent space $\{(\theta, \eta)\}$, where θ is referred to as the target ability (the unidimensional ability the test is designed to measure) and η is referred to as a nuisance ability (the ability being measured by the test that is not part of the target ability, η possibly multidimensional).

Denote $P(\theta, \eta)$ as the joint IRF for the studied item. The combined (θ, η) is assumed complete. Hence it follows that $P(\theta, \eta)$ does not differ across groups. The marginal IRFs of the studied item for the focal and reference groups can be obtained by the following integrations

$$E_F[Y|\theta] \equiv E_F[P(\theta, \eta)|\theta] \equiv \int P(\theta, \eta) f_F(\eta|\theta) d\eta,$$

and

$$E_R[Y|\theta] \equiv E_R[P(\theta, \eta)|\theta] \equiv \int P(\theta, \eta) f_R(\eta|\theta) d\eta,$$

where $f_F(\eta|\theta)$ and $f_R(\eta|\theta)$ are the conditional density functions of η at a given θ for the focal and reference groups, respectively. Here the IRFs are **marginal** in the sense that η is integrated out according to the conditional densities. Hence, latent variable DIF is said to take place at a given θ level if $E_R[Y|\theta] \neq E_F[Y|\theta]$ (agreeing with Definition 2), and the amount of DIF at θ is measured by

$$B_0(\theta) \equiv E_R[Y|\theta] - E_F[Y|\theta]. \tag{4}$$

It is clear from the discussions in the previous section that an IRT-based DIF definition for ordinally scored polytomous items can be given from the item structure invariance perspective:

Definition 3 *(Latent Variable IRT Version of DIF) DIF is said to occur at θ if $B_0(\theta) \neq 0$.*

A global index of DIF in the ordinal polytomous case can be defined in exactly the same way as Shealy and Stout (1993) do for the dichotomous case:

$$\beta = \int B_0(\theta) f_F(\theta) d\theta, \tag{5}$$

that is, as the expected amount of DIF experienced by a randomly selected F group examinee.

Extension of SIBTEST to Polytomous Case

Designed for use in in the dichotomous case, SIBTEST is a statistical procedure for testing $H_0 : \beta = 0 \quad vs. \quad H_A : \beta \neq 0$. A statistical test procedure for detecting polytomous DIF is obtained by some minor modifications to SIBTEST. The modified procedure includes dichotomous item scoring as a special case. In using the modified SIBTEST, the studied item can be either dichotomous or polytomous, and the matching subtest can consist of both types of items. Hence, the modified SIBTEST is a generalized version of SIBTEST. In the following we will give a brief description of the modified SIBTEST procedure (see Shealy & Stout, 1993, for a full description of the original SIBTEST procedure). Note that dichotomous items are special cases of polytomous items. Readers can easily verify that all the notations and results in the following apply to both dichotomous and polytomous items.

Preliminary Notation:

Y: Studied item score which has $m + 1$ categories (i.e., $Y = 0, 1, ..., m$);

$X_1, X_2, ..., X_n$: item scores for n matching items;

$m_1, m_2, .., m_n$: the maximum possible scores for $X_1, X_2, ..., X_n$, respectively, e.g., $X_i = 0, 1, ... m_i$; if $m_i = 1$, then X_i is dichotomously scored;

$X = \sum_{j=1}^{n} X_j$: matching score; $X = 0, 1, ..., n_H$, where

$$n_H \equiv \sum_{j=1}^{n} m_j \tag{6}$$

is the maximum possible matching score;

$\bar{Y}_{gk}$: average score on studied item for all group g ($g = F$ or R) examinees for which $X = k$;

p_{Fk}: proportion of examinees in the focal group getting score $X = k$ on the matching subtest $X_1, ..., X_n$; i.e., $p_{Fk} = N_{Fk}/N_F$, where N_F is the total number of focal group examinees and N_{Fk} is the total number of focal group examinees with matching test score $X = k$;

p_k: propoιtion of examinees getting score $X = k$ on the matching subtest $X_1, ..., X_n$, i.e., $p_k = N_k/N$, where N is the number of total examinees and N_k is defined by

$N_k = N_{Rk} + N_{Fk}$.

It should be indicated here that only two modifications to the SIBTEST procedure are needed in order to make it applicable to polytomously scored items. One is to replace n in the SIBTEST test statistic (Shealy & Stout, 1993) with n_H as defined in (6) so that there will be $n_H + 1$ matching scores; and the other is to replace the KR20 coefficient α estimate used by Shealy and Stout in their true score regression estimators (see (21) and (22) below) with Cronbach's standard coefficient α estimate (Lord & Novick. p.89. 1968):

$$\hat{\alpha} = \frac{n}{n-1}\left[1 - \frac{\sum_{j=1}^{n} \hat{\sigma}_{X_j}^2}{\hat{\sigma}_X^2}\right], \tag{7}$$

where $\hat{\sigma}_{X_j}^2$ is the sample variance of X_j; $j = 1, ..., n$; and, $\hat{\sigma}_X^2$ is the sample variance of $X = \sum_{j=1}^{n} X_j$. Notice that the KR20 formula is for dichotomous items only, while formula (7) can be generally used for either polytomous or dichotomous items. In the dichotomous case, the two expressions are identical.

Understanding SIBTEST

Since the adaptation of SIBTEST to the polytomous case is so straightforward, an important role of Section 3 is in presenting a clear explanation of the SIBTEST procedure. Despite its IRT origin, SIBTEST can also be understood from a classical test theory perspective and, as Shealy and Stout (1993) point out, its use is not predicated on adopting their multidimensional IRT-based conception of DIF. The procedure neither requires nor uses IRT ability or item parameter estimates for its calculation.

Assume a classical test (true score) theory representation for X, i.e., $X = T + E$, where T denotes the matching test true score for a randomly selected examinee. The variable $E = X - T$ denotes the measurement error and is assumed to have mean zero in both groups. Let $f_g(t)$ be the density of matching test true scores in group g, $g = R$ or F. We will consider the regression of studied item score on matching test true score in group g, and denote this regression as $E_g[Y|t] = E[Y|T = t, G = g]$.

Definition 4 *(Latent Variable Version of Null-DIF via True Score Theory) A studied item is defined as being free of DIF when, for all values of true matching score t, $E_R[Y|t] = E_F[Y|t]$.*

If matching test items are ordinally scored polytomous items and follow the partial credit model, generalized partial credit model, or the logistic graded response model, matching test true scores are a strictly monotonic transformation of θ (Chang, 1994). Therefore, matching-test true scores are simply a one-to-one transformation of the latent variable θ. According to Chang and Mazzeo (1994), an item being free of DIF according to Definition 4 implies that the item parameters in $E_R[Y|t]$ must be the same as those in $E_F[Y|t]$. Such a null DIF definition as Definition 4 is consistent with the definition of measurement invariance given in Meredith and Millsap (1992). Under the above conditions, Definition 4 is equivalent to an IRT latent variable definition formulated in terms of identity of item parameters for all ICRFs across groups. This result is important in that it indicates that no information is lost by comparing only conditional item score means in the assessment of IRT DIF. Note that in the dichotomous case, Definitions 2 and 4 are equivalent.

The SIBTEST procedure was designed primarily to detect what Shealy and Stout (1993) refer to as unidirectional DIF.

Definition 5 *(Unidirectional DIF, a latent variable true score theory version) A studied item exhibits unidirectional DIF against the focal group if, for all t, $E_R[Y|t] \geq E_F[Y|t]$, and $E_R[Y|t] > E_F[Y|t]$ for some values of t; a studied item exhibits unidirectional DIF against the reference group if, for all t, $E_R[Y|t] \leq E_F[Y|t]$, and $E_R[Y|t] < E_F[Y|t]$ for some values of t.*

Analogous to (4), the local measure of DIF at t can be defined as $B(t) = E_R[Y|t] - E_F[Y|t]$. Thus, the corresponding DIF index with respect to a density $f_F(t)$ can be defined as $\beta = \int B(t) f_F(t) dt$. In the unidimensional IRT case, β is of course the same as in (5), as seen by a change of variable calculus argument.

By naive intuition, DIF could be estimated locally by the values of

$$d_k = \bar{Y}_{Rk} - \bar{Y}_{Fk}, \quad k = 0, \ldots, n_H, \tag{8}$$

because $\bar{Y}_{Rk} - \bar{Y}_{Fk}$ is the group difference in performance on the studied item among examinees with the same matching test score. This is consistent with the observed-score definition of null-DIF. If we are able to validly assume examinees with the same matching test score have (approximately) the same θ ability (which in fact will be true if either the matching test is long enough to be highly reliable or the groups under study have similar θ ability distributions), then (8) is also approximately the difference in item scores at the same target ability. If the studied item does not have observed-score DIF, we expect $d_k \approx 0$. Thus, a suggested statistic to estimate the DIF β for the special case where R and F have the same target θ distribution is:

$$\hat{\beta} = \sum_{k=0}^{n_H} p_k d_k. \tag{9}$$

Note that, instead of p_{Fk}, weights p_k here are used for better Type I error control (see Shealy & Stout, 1993). A test statistic can be defined by

$$B = \frac{\hat{\beta}}{\hat{\sigma}(\hat{\beta})}, \tag{10}$$

where

$$\hat{\sigma}(\hat{\beta}) = \left[\sum_{k=0}^{n_H} {p_k}^2 \left(\frac{\hat{\sigma}^2(Y|k,R)}{N_{Rk}} + \frac{\hat{\sigma}^2(Y|k,F)}{N_{Fk}} \right) \right]^{1/2},$$

and $\hat{\sigma}^2(Y|k,g)$ is the sample variance of the studied item scores for examinees in group g ($g = R$ or F) with a total score $X = k$ on the matching subtest. When R and F have the same target ability distributions, it is not difficult to justify by the central limit theorem that, when $\beta = 0$, the distribution of B defined in (10) is $\approx N(0, 1)$.

Regression Correction–Type I Error Control. Unfortunately, the test statistic B defined in (10) tends to display highly inflated impact-induced (i.e., caused by the existence of group ability differences) Type I error rate. To further explain this, let the studied item have no DIF, and let reference matching true score T_R be stochastically greater than that for focal group, say, T_F. It is obvious that $\bar{Y}_{Rk}$ and $\bar{Y}_{Fk}$ are unbiased estimators of $E[Y|X = k, G = R]$ and $E[Y|X = k, G = F]$ respectively. By the result from elementary statistics that $E[E[A|B]] = E[A]$ for random variables A and B, we obtain

$$E[\bar{Y}_{Rk}] = E[Y|X = k, G = R] = E[I_R(T_R)|X = k, G = R], \tag{11}$$

where $I_R(T_R) = E[Y|T_R, X = k, G = R]$. Similarly,

$$E[\bar{Y}_{Fk}] = E[Y|X = k, G = F] = E[I_F(T_F)|X = k, G = F], \tag{12}$$

where $I_F(T_F) = E[Y|T_F, X = k, G = F]$. Note that I_R and I_F are item response functions for the studied item for R and F respectively. Since the studied item and the matching items have no DIF, it is easy to see that

$$I_R(t) = I_F(t), \quad for\ all\ values\ of\ t. \tag{13}$$

Because T_R is stochastically greater than T_F, it is also clear that the distribution of T_R conditioning on $X = k$ is stochastically larger than the distribution of T_F conditioning on $X = k$. Then by (11), (12), and (13), and noting that $I_R(t) = I_F(t)$ can be shown to be increasing in t by the assumption of monotonicity of IRFs, we expect

$$E[d_k] = E[\bar{Y}_{Rk}] - E[\bar{Y}_{Fk}] > 0 \tag{14}$$

for all k. Thus, matching on matching test observed score does not produce true score matching. Let $t_{R,k}$ denote the expected matching test true score for reference group examinees with a matching test score of k, i.e.,

$$t_{R,k} = E[T|X = k, G = R], \tag{15}$$

and let $t_{F,k}$ denote the analogous quantity for the focal group. Noting (8) and (9), and using (14), in general, the B of (10) is statistically inflated when the focal and the reference groups have different distributions of θ, even if no DIF is present. In particular, when the focal group true-score distribution is stochastically less than that of the reference group, one might expect that the d_k will tend to indicate spurious DIF by confounding impact, the difference between $t_{R,k}$ and $t_{F,k}$, with DIF, the difference in $E_R[Y|t]$ and $E_F[Y|t]$.

Just as with SIBTEST in the dichotomous case, the modified SIBTEST estimates

$$E_R[Y|t_k] - E_F[Y|t_k], \quad k = 0, ..., n_H \tag{16}$$

where $t_k = (t_{R,k} + t_{F,k})/2$. Note that t_k in (16) is the same for R and F groups. Let $S_g(t) = E_g[Y|t]$. Then (16) can be rewritten as

$$S_R(t_k) - S_F(t_k), \quad k = 0, 1, ..., n_H. \tag{17}$$

Since (17) is the difference in expected studied item scores for the two groups where the examinees are matched according to the same matching test true score t_k, $k = 0, 1, ..., n_H$. if there is no latent variable DIF, examinees with the same matching true score will have the same expected studied scores, regardless of group membership.

Now we have to estimate the terms of (17). Assume that $S_g(t)$ is a locally linear function of t. That is, intuitively, it is allowed curvature, but no abrupt change in slope. By Taylor expansion. $S_R(t) \approx S_R(t_0) + S_R{}'(t_0)(t - t_0)$, where $S_R{}'(t_0)$ can be obtained by mean value theorem

$$S_R{}'(t_0) \approx \frac{S_R(b) - S_R(a)}{b - a},$$

where $a \leq t_0 \leq b$. Thus,

$$S_R(t_k) \approx S_R(t_{R,k}) + S_R{}'(t_{R,k})(t_k - t_{R,k}). \tag{18}$$

It is obvious that $t_{R,k-1} \leq t_{R,k} \leq t_{R,k+1}$. Also

$$S_R{}'(t_{R,k}) \approx \frac{S_R(t_{R,k+1}) - S_R(t_{R,k-1})}{t_{R,k+1} - t_{R,k-1}},$$

where $S_R(t_{R,k+1})$, $S_R(t_{R,k})$, and $S_R(t_{R,k-1})$ can be estimated by $\bar{Y}_{R,k+1}$, $\bar{Y}_{Rk}$, and $\bar{Y}_{R,k-1}$, respectively. Following Shealy and Stout (1993), as (18) suggests we estimate $S_R(t_k)$ by

$$\bar{Y}^*_{Rk} = \bar{Y}_{Rk} + \frac{\bar{Y}_{R,k+1} - \bar{Y}_{R,k-1}}{t_{R,k+1} - t_{R,k-1}}(t_k - t_{R,k}). \tag{19}$$

Similarly, $S_F(t_k)$ can be estimated by

$$\bar{Y}^*_{Fk} = \bar{Y}_{Fk} + \frac{\bar{Y}_{F,k+1} - \bar{Y}_{F,k-1}}{t_{F,k+1} - t_{F,k-1}}(t_k - t_{F,k}). \tag{20}$$

In order to compute (19) and (20), one has to obtain the true score regression estimates:

$$\hat{t}_{R,k} \equiv \hat{E}[T|X = k, G = R], \tag{21}$$

$$\hat{t}_{F,k} \equiv \hat{E}[T|X = k, G = F], \tag{22}$$

taking

$$\hat{t}_k = \frac{\hat{t}_{R,k} + \hat{t}_{F,k}}{2}. \tag{23}$$

These true score regression estimates (21) and (22) are obtained by assuming a linear regression approximation to the matching test true score regression on matching test observed score and estimating its slope using a classical coefficient α reliability estimate (see (7)).

Define $d_k^* = \bar{Y}_{Rk}^* - \bar{Y}_{Fk}^*$, $k = 0, ..., n_H$. Equations (19) - (23) imply that d_k^* is obtained by a transformation of $\bar{Y}_{Rk} - \bar{Y}_{Fk}$, i.e.,

$$\bar{Y}_{Rk} - \bar{Y}_{Fk} \rightarrow \bar{Y}_{Rk}^* - \bar{Y}_{Fk}^*. \tag{24}$$

This transformation is referred to as "*the regression correction*" by Shealy and Stout (1993). Therefore, instead of $\hat{\beta}$ in (10), $\hat{\beta}^* = \sum_{k=0}^{n_H} p_k d_k^*$ is employed in the modified SIBTEST and as well as in the original SIBTEST. Notice that when (14) happens as a result of impact, (24) removes the statistically inflating effect of impact so that the distribution of $\bar{Y}_{Rk}^* - \bar{Y}_{Fk}^*$ has approximately mean 0 when H_0 is true (see Shealy & Stout, 1993, for details). Moreover, B is $\approx$ N(0,1) when H_0 is true. Thus, we reject H_0 if $|B| > 1.96$ at significance level 0.05 under the alternative hypothesis of DIF against either group.

Two Related Procedures: Mantel and SMD.

It should be mentioned that when the studied item is included in the calculation of matching test score, B in (10) is just the SMD index. The SMD procedure is a generalization of the standardization method (Dorans & Kulick, 1986). The Mantel procedure is an extension of the MH procedure and is specifically generalized for ordered categorical data. It is very similar to the SMD procedure in that it compares the average studied item score conditioning on the matching score for the R and F groups. The Mantel procedure (as adapted for use in DIF research) also includes the studied item in its matching test.

The fundamental difference between the SIBTEST and the Mantel and SMD procedures is the way they confront the statistical inference problem caused by the presence of a difference between focal and reference target ability distributions. The Mantel and SMD procedures deal with this problem by including the single studied item in the matching test, while SIBTEST controls such inflated Type I error by the "regression correction", a non-parametric model based linear transformation.

Results from the Mantel and SMD procedures will be compared with that of the

SIBTEST procedure in a section below. (See Zwick et. al., 1993, and Dorans & Schimitt, 1993, for details about these procedures.)

4. Simulation Studies

Two simulation studies are included in this section. The first one compares SIBTEST with the Mantel and SMD procedures for data simulated under the Rasch-like model constraints of equal discriminations across polytomous items like those imposed by the partial credit model (Masters, 1982). The second one compares SIBTEST performance with that of the Mantel and SMD procedures with "non-Rasch" data, as generated by the generalized partial credit model (Muraki, 1992), which allows varying discrimination across polytomous items. The first study indicates that, in terms of Type-I error rate and power, the performance of SIBTEST is essentially as good as that of the Mantel and SMD procedures for data constrained by equal discrimination across polytomous items. The second study indicates that, when the equal discrimination constraint is violated in a realistic manner, the Mantel and SMD procedures each tends to have unacceptably high rates of either Type-I or Type-II errors. More specifically, the Mantel and SMD procedures tend to flag a non-DIF item as a DIF item when it is either more or less discriminating than the average discrimination of the items that make up the matching test. Moreover, the Mantel and SMD procedures often fail to flag a DIF item if it is less discriminating than the average discrimination of the matching test. By contrast, the modified SIBTEST procedure was found to be much more robust with regard to controlling impact-induced Type I/Type II error rates inflation than the other procedures.

Study I: Polytomous Items that Follow Partial Credit Model

Zwick et al. (1993) conducted a simulation study to evaluate the Mantel procedure in detecting DIF with polytomous items following the partial credit model (Masters, 1982). Their design included 54 conditions across different items, ability groups, and types of DIF. The study is so well designed that we resolved to use the same design to compare the performance of the SIBTEST procedure with the Mantel and SMD procedures for polytomous items which follow the Rasch-like partial credit model.

Matching Test. The matching subtest consisted of 24 items (25 for Mantel and SMD, because of inclusion of studied item), 20 dichotomously scored items and 4 (5 for Mantel and SMD) 4-category ordinally-scored items (scored as 0, 1, 2, or 3). The 3PL model defined below with a common c=0.15 was used to generate item responses for the dichotomous items.

$$P_j(\theta) = c + \frac{1-c}{1+\exp\{1.7a_j(\theta - b_j)\}}, \quad j = 1, \ldots, 20. \tag{25}$$

The partial credit model defined below was used for generating item responses for the 4-category ordinally-scored items.

$$P_{jk}(\theta) = \frac{\exp\{\sum_{i=0}^{k}(\theta - b_{ji})\}}{\sum_{i=0}^{3}\exp\{\sum_{i=0}^{l}(\theta - b_{ji})\}}, \qquad k = 0, 1, 2, 3, \tag{26}$$

where $j = 21, 22, 23, 24$ indicates item number, and for notational convenience, $\sum_{i=0}^{0}(\theta - b_{ji}) \equiv 0$. It is to be noted that the partial credit model is a natural generalization of the Rasch dichotomous model in that all items have the same discrimination parameters. It should also be noted that while the usual scale factor of D=1.7 was used in the 3PL model (25) for the matching test, this factor is not employed in (26). Thus the model in (26) cab be equivalently written by including a discrimination parameter of 0.588 along with a scaling factor D=1.7 is used. The item parameters for the matching test items are identical to those used by Zwick et. al.. (See Table 1.)

DIF and Null DIF Modeling. The three studied items are all 4-category polytomous items. The partial credit model defined in (26) was used to create DIF by adding a group index g to the item parameters in order to simulate responses of studied items for focal and reference groups respectively. Specifically,

$$P_{jkg}(\theta) = \frac{\exp\{\sum_{i=0}^{k}(\theta - b_{jig})\}}{\sum_{l=0}^{3}\exp\{\sum_{i=0}^{l}(\theta - b_{jig})\}}, \qquad k = 0, 1, 2, 3, \tag{27}$$

Dichotomous Item Parameters			
Item Number	a_i	b_i	c_i
1	0.741	-2.25	0.15
2	0.861	-2.00	0.15
3	1.162	-1.75	0.15
4	0.638	-1.50	0.15
5	1.000	-1.25	0.15
6	1.000	-1.00	0.15
7	1.162	-0.75	0.15
8	0.638	-0.50	0.15
9	0.741	-0.25	0.15
10	0.861	0.00	0.15
11	1.000	0.00	0.15
12	0.741	0.25	0.15
13	1.162	0.50	0.15
14	0.638	0.75	0.15
15	0.861	1.00	0.15
16	0.638	1.25	0.15
17	0.741	1.50	0.15
18	1.162	1.75	0.15
19	1.000	2.00	0.15
20	0.861	2.25	0.15
Polytomous Item Parameters			
Item Number	b_{i1}	b_{i2}	b_{i3}
21	-0.91	-0.93	1.29
22	-1.34	1.72	3.40
23	-1.76	0.09	0.19
24	-2.20	-1.33	-0.48

Table 1: Item Parameters in the Matching Test (Items 1-24)

where $j = 1, 2, 3$ indicates studied item number and $g = R, F$ denotes reference or focal group. For each studied item, the four types of DIF were modeled. The DIF types were referred to by Zwick et. al. (1993) as *constant, low-shift, high-shift,* and *balanced DIF.* Figure 1 shows each type of DIF in terms of IRFs, for one of the three studied items. It is clear from the figure that the first three types DIF are unidirectional and the fourth is non-unidirectional. For the three studied items, the reference group threshold parameters were (-.91,.98,.21), (-2.25,-1.8,1.66), and (-.54,-2.11,.74), respectively. The focal group studied item threshold parameters were defined by $b_{jiF} = b_{jiR} \pm C$, where $C = 0.0$, 0.1 or 0.25. The $C = 0$ condition results in a null DIF condition. Each of the four types of DIF were crossed with 2 nonzero magnitudes of DIF ($C =.10$ and $C =.25$), yielding a total of 8 DIF conditions each containing 3 studied items. Thus, the simulation involved the use of 27 studied items, 24 of which exhibited DIF of varying types and to varying degrees and three of which exhibited no DIF. (See Zwick et. al. (1993) for detailed descriptions about the DIF modeling.)

INSERT FIG 1 ABOUT HERE

Reference and Focal Groups. The performance of the DIF detection procedures was studied under two conditions: (1) no difference in ability distributions across groups; and (2) a substantial but realistic difference in ability distributions between the two groups. For the reference group, θs were always sampled from N(0,1). For the focal group, however, two conditions were considered, N(0,1) and N(-1,1). For each replication, 500 θ values were sampled for each of the two groups.

Simulation Procedure. There were a total of 54 cells in the design, 2 focal group θ distributions (N(0,1) and N(-1,1)) crossed with 27 studied items (4 types of DIF crossed with two magnitudes of DIF crossed with 3 DIF items plus 3 null DIF items). For each cell, 600 replications were carried out. Each replication involved the following steps: (1) generate 500 θ values for reference and focal groups respectively, according to appropriate distributions; (2) generate item responses according to the appropriate item parameters for each sampled θ value; (3) perform the three procedures to calculate their respective DIF statistics.

Simulation Results of Study I

Table 2 provides rejection rates for all nine conditions. Each percentage (in the Average columns) in the table is averaged over the three studied items and the two distribution conditions. There are 600 replications for each studied item and each distribution condition, so that the proportions in the table (in the Average columns) are based on 3600

Type of DIF	Mantel	SMD		SIBTEST	
	Rej. Rate	Rej. Rate		Rej. Rate	
	Average	Average	Average	Focal:N(0,1)	Focal:N(-1,1)
NULL	0.049	0.046	0.063	0.041	0.085
Constant 0.10	0.176	0.174	0.173	0.189	0.156
Constant 0.25	0.733	0.720	0.698	0.751	0.644
Balanced 0.10	0.045	0.045	0.061	0.045	0.078
Balanced 0.25	0.061	0.059	0.068	0.067	0.069
Low Shift 0.10	0.063	0.062	0.071	0.053	0.089
Low Shift 0.25	0.125	0.119	0.119	0.103	0.135
High Shift 0.10	0.058	0.058	0.064	0.063	0.065
High Shift 0.25	0.121	0.127	0.126	0.149	0.102

Table 2: Rejection Rates for Mantel, SMD, and SIBTEST Procedures.

replications.

The type I error rates were 0.049 and 0.046 for the Mantel and SMD procedures respectively, excellent adherence to the nominal Type I error rate. By comparison, the Type I error rate for the SIBTEST procedure was 0.063 (0.041 for the focal condition N(0,1), and 0.085 for the focal condition N(-1,1), recalling that the reference condition is always N(0,1)), a slight inflation.

Besides Type I error, Table 2 also presents the rejection rates in the power study. SIBTEST appears to be as powerful as Mantel and SMD in detecting unidirectional DIF items. For the constant 0.10 DIF condition, rejection rates were 0.176 and 0.174 for the Mantel and SMD procedures, respectively. For the SIBTEST procedure, the corresponding rate was 0.173. For the constant 0.25 DIF condition, rejection rates were 0.733 and 0.720 for the Mantel and SMD procedures, respectively. For the SIBTEST procedure, the corresponding rate was 0.698 (0.751 for the focal condition N(0,1) and 0.644 for the focal condition N(-1,1)). Thus, the SIBTEST procedure had a slightly lower rejection rate than the Mantel and SMD procedures. When the focal and reference distributions are different, the SIBTEST procedure tends to have either slightly inflated Type I error rate or slightly lower power. These phenomena may be due to a tendency to "*over regression-correct*" for group ability differences. In other words, SIBTEST seems to push the center of the testing statistic's distribution from the right hand side of 0 before the regression correction slightly to the left hand side of 0 after the regression correction. It should be noted that the results of Mantel reported here are very similar to the those reported in Zwick et al.

Study II: Polytomous Items that Follow a Generalized Partial Credit Model

Study I suggests that the performance of the Mantel and SMD procedures was slightly better than that of SIBTEST when used with polytomous items that follow the Rasch-like partial credit model, although the SIBTEST performance is good nonetheless. The focus of Study II is on the performance of these three DIF procedures for polytomous items that vary in discrimination. Such a study is important from an applications perspective to determine for matching tests that are moderate in length whether DIF procedures produce unacceptably high Type I error rates for studied items with realistic item parameters that are more or less discriminating than the items that make up the matching test.

The design of Study II is similar to the dichotomous item DIF study reported by Roussos and Stout (1994). Their study clearly shows that the Mantel Haenszel procedure has much greater Type I error inflation rates than does SIBTEST when studied-item discrimination parameters depart from those included in the matching test. By analogy, one would suspect that the Mantel/SIBTEST comparisons might produce similar results in the polytomous case.

Matching Test. A single 24-item (25 for Mantel and SMD) matching test was used in Study II for SIBTEST. The matching test consisted of 20 dichotomous items and 4 (5 for Mantel and SMD) 4-category polytomous items. The dichotomous items were assumed to follow a 3PL model and the same item parameters as in Study I (see Table 1) were used. For matching test and studied polytomous items, however, we used a generalized partial credit model (Muraki, 1992):

$$P\{X_j = k|\theta\} = \frac{\exp\{1.7a_j \sum_{v=0}^{k}(\theta - b_{jv})\}}{\sum_{v=0}^{3}\exp\{1.7a_j \sum_{c=0}^{v}(\theta - b_{jc})\}} \qquad k = 0, 1, 2, 3. \qquad (28)$$

Polytomous Item Parameters				
Item Number	a_i	b_{i1}	b_{i2}	b_{i3}
21	0.563	-2.449	-0.089	2.416
22	1.359	-0.342	1.008	1.797
23	0.535	-1.804	-0.368	0.219
24	0.779	-0.345	2.428	2.822

Table 3: Item Parameters for Polytomous Items in Study II Matching Test (Items 21-24)

where a_j denotes the discrimination parameter, and $\sum_{v=0}^{0}(\theta - b_{jv}) \equiv 0$ for notational convenience. It should be noted that the partial credit model defined in (26) is a special case of the generalized partial credit model with $a_j \equiv 0.588$. Obviously, the equation in (28) can be easily used to represent items with different discriminations by varying the value of the a_j-parameter. Table 3 contains the item parameters for the matching test polytomous items. These parameter values were taken from an actual calibration for the 1992 National Assessment of Educational Progress (Johnson and Carlson, 1994). The resulting matching test has an average a_j value of 0.869, and the standard deviation of a_j is 0.218.

Studied Items. A total of eleven polytomous studied items were included in Study II. These items shared a common set of reference group threshold parameters (-1,0,1) but had different discrimination parameters. The eleven values for the discrimination parameters were 2.0, 1.5, 1.36, 1.12, 1.0, 0.869, 0.588, 0.33, 0.25, 0.23, and 0.15. A strong case for the appropriateness and the realism of the selected item discrimination parameters can be made. The value 0.869 is the average matching test discrimination. The other values were chosen based on the actual calibration of the 1992 NAEP analysis referred to above. (Johnson and Carlson, Appendix E, 1994). The values 0.33 and 1.12 are respectively the 10-th and 90-th percentiles of the distribution of discrimination parameter estimates for the 1992 NAEP Reading item pool. The 0.23 and 1.36 values are respectively the smallest and largest estimated discrimination parameters in the pool. It is certainly possible that other tests could produce more extreme values. Thus, the range of discrimination parameters from 0.25 to 1.5 for the simulation was extruded still further to include the values 0.15 and 2.0.

Two DIF conditions were examined. The first condition was the null-DIF condition in which the R and F groups studied-item parameters were identical. The second condition was the same "constant" DIF condition in Study I. The level of threshold shift was 0.25. Thus, the focal group thresholds for the studied items were (-0.75,0.25,1.25) compared to (-1,0,1) for the reference group. The threshold shift of 0.25 was paired with each of the 11 discrimination parameters. It should be noted however, that the amount of DIF in the studied item (as measured in terms of differences between IRFs) depends on both the threshold and the discrimination parameters values.

Ability Distributions. Previous work with a variety of DIF procedures suggests that many work well when there are no group ability differences. Therefore, Study II considered only the case where such differences are present. Focal group abilities were sampled from N(-1,1). Reference group abilities were sampled from N(0,1). Thus, in the second study, the F group ability distribution is stochastically smaller than the R group distribution. Several researchers (e.g., Mullis, Dossey, Owen, & Phillips, 1993, and Donoghue, Holland, & Thayer, 1993) report that a difference in ability means of one standard deviation is often common between certain focal and reference groups of interest. Thus the choice of ability distributions is realistic.

Simulation procedure. The single matching test was paired with each of the 11 null-DIF studied items and each of the 11 DIF magnitudes to form 22 experimental conditions. For each condition, 1,000 replications were carried out. For each replication, 500 simulated θ values were sampled from the focal and reference group distributions. For five of the 11 Null DIF conditions (a=.23, .33, .588, 1.12, 1.36) additional replications were carried out at increased sample sizes. An additional 1,000 replications of the procedures were run with both R and F sample sizes set at 1,000 and 2,000. For each simulee, item responses to matching test and studied items were generated in accordance with the model described above. The simulated item responses were then analyzed with each of the three DIF procedures.

Simulation Results of Study II

Type I Error Rate Study. Table 4 provides Type I error rates for the 11 null-DIF items for the case of a sample size of 500 for each group. Both the Mantel and SMD procedures exhibit highly inflated Type I error rates (with rejection rates as high as 40%) when the discrimination parameter of the studied item differs from the average discrimination of the matching test items, i.e., $a_j = 0.869$. A somewhat surprising result is that the studied item for which Mantel and SMD exhibited error rates closest to the nominal 0.05 level (i.e., $a_j = 0.588$) was not the studied item with a parameter value equal to that average matching test discrimination. It may be that both average level and degree of dispersion of matching test discriminations are important or that the relationship is complicated by the presence of non-zero guessing parameters for the dichotomous items. This is an interesting topic for further research. Note also that, when $a = 0.588$, the generalized partial credit model in (28) is equivalent to the partial credit model in (26). By contrast, SIBTEST, while displaying evidence of mild increased error rates for more extreme discrimination parameters, is considerably more robust. Its Type 1 error rates ranged only from 0.061 to 0.093 across the eleven studied items.

Table 5 presents Type I error rates for the selected 5 null-DIF items that were run with expanded sample sizes. It is evident that the Type I error rates of Mantel and SMD are driven not only by the discrimination parameter values but also by sample sizes. When the sample size increases for both R and F groups within a realistic range, the false rejection rates increase dramatically. It marked contrast that the Type I error rate of the SIBTEST procedure remains reasonably consistent despite the increase in sample sizes (from 500/500 to 1000/1000 and even to 2000/2000).

In hypothesis testing, it is often true (as it should be) that *power* will increase when sample size increases. By contrast, Table 5 shows that the *Type I error rates* of Mantel

a_j-Parameters	Mantel	SMD	SIBTEST
2.000	0.450	0.415	0.093
1.500	0.310	0.303	0.082
1.360	0.236	0.222	0.078
1.120	0.154	0.148	0.078
1.000	0.114	0.101	0.077
0.869	0.098	0.091	0.079
0.588	0.053	0.049	0.070
0.330	0.218	0.229	0.073
0.250	0.239	0.240	0.083
0.230	0.282	0.291	0.061
0.150	0.417	0.405	0.086

Table 4: Empirical Type I Error for Three Polytomous DIF Procedures as a Function of Studied Item Discrimination Parameter. (Results are based on 1,000 replications and a sample size of 500 for each group.)

Sample Size: 500/500			
a-values	Mantel	SMD	SIBTEST
a=1.360	0.236	0.222	0.078
a=1.120	0.154	0.148	0.078
a=0.588	0.053	0.049	0.070
a=0.330	0.218	0.229	0.073
a=0.230	0.282	0.291	0.061

Sample Size: 1000/1000			
a-values	Mantel	SMD	SIBTEST
a=1.360	0.438	0.404	0.081
a=1.120	0.311	0.286	0.074
a=0.588	0.048	0.051	0.081
a=0.330	0.227	0.229	0.065
a=0.230	0.446	0.437	0.062

Sample Size: 2000/2000			
a-values	Mantel	SMD	SIBTEST
a=1.360	0.723	0.702	0.108
a=1.120	0.484	0.463	0.088
a=0.588	0.058	0.063	0.081
a=0.330	0.411	0.394	0.063
a=0.230	0.747	0.742	0.064

Table 5: Empirical Type I Error Rates for Three Polytomous DIF Procedures as a Function of Reference/Focal Sample Size. (Results are based on 1,000 replications. Rates for the 500/500 sample size are reproduced from Table 4.)

a_j-Parameters	Mantel	SMD	SIBTEST
2.000	1.000	1.000	0.985
1.500	1.000	0.999	0.962
1.360	1.000	0.999	0.953
1.200	0.995	0.992	0.891
1.000	0.982	0.979	0.857
0.869	0.961	0.953	0.811
0.588	0.721	0.679	0.637
0.330	0.127	0.118	0.353
0.250	0.042	0.043	0.243
0.230	0.056	0.054	0.223
0.150	0.131	0.136	0.138

Table 6: Rejection Rates for Three Polytomous DIF Procedures as a Function of Studied Item Discrimination Parameter for Studied Items Displaying Constant DIF. (Results are based on 1,000 replications.)

and SMD increase as the sample size increases. This result may at first seem implausible. On further reflection however the pattern of results is quite sensible. In describing the Mantel and SMD Null-DIF rejection rates, the term "Type I error rate" is somewhat misleading. The null hypothesis of Mantel and SMD is not equivalent to the IRT-based null-DIF definition used in the simulation.

In order to explain this more clearly, it is helpful to repeat the three null DIF definitions introduced earlier in the paper.

$$\begin{aligned} \textit{Observed score } H_0: E_R[Y|X] &= E_F[Y|X] \textit{ for all values of} X. &(29) \\ \textit{Latent variable true score } H_0: E_R[Y|t] &= E_F[Y|t] \textit{ for all values of } t. &(30) \\ \textit{Latent variable IRT } H_0: E_R[Y|\theta] &= E_F[Y|\theta] \textit{ for all values of } \theta. &(31) \end{aligned}$$

For theoretical and practical reasons, the latent variable definitions are apparently preferable. Most Monte Carlo studies (e.g. Donoghue et. al., 1993, Welch & Hoover, 1993, Zwick et. al., 1993) and theoretical papers (e.g. Holland & Thayer, 1988, Meredith & Millsap, 1992, Zwick, 1990) use the definitions and methods associated latent variable DIF in modeling and simulating DIF.

Since the studied items were the same for the for both R and F groups, the data simulated has no DIF and thus (31) is true. Since the null hypothesis described by (30) is equivalent to (31), the rejection rates reported in Table 5 for the SIBTEST procedure are really *Type I error rates* for items that are free of DIF. However, the observed score null hypothesis paraphrased in (29) that in effect is being tested by the Mantel and SMD procedures is not true under this simulation design (or, for that matter in most DIF simulation studies which use IRT definitions and methods to model DIF and simulated data). In general, (29) does not coincide with (31) when (i) the ability distribution of one group is stochastically larger than that of the other; and (ii) the studied item is more or less discriminating than the average item discrimination in the matching test (Meredith & Millsap, 1992, Zwick, 1990). Therefore, except for the $a = 0.588$ case, the "Type I error rate" reported in Table 5 for Mantel and SMD should from the statistical perspective be renamed as "*power*", i.e., the probability of rejecting the null hypothesis (29) when it is not true. That is, the increase of "Type I error" is merely the expected increase in power as sample size increases. It does not seem wise to rename the term here, however. Obviously, the observed score based Null-DIF definition (29) should be used with great caution, if one really wants to distinguish impact from latent variable DIF. For that matter, the latent variable definitions are apparently preferable.

Power Study. Table 6 presents rejection rates for the 11 DIF items. All three procedures exhibited higher rejection rates for the items with larger discrimination parameters. This is as expected since, for a fixed shift in thresholds, the amount of DIF in the item is directly dependent on the discrimination parameter. For the seven largest discrimination parameters, the Mantel and SMD procedures did exhibit higher rejection rates than did SIBTEST, although the SIBTEST rates were quite good nonetheless. Except for the $a_j = 0.588$ case, the greater Mantel and SMD rejection rates are not caused by they having greater power against the latent variable null-DIF hypotheses but indeed is explained by differences in Type I error rates. That is, the power of any hypothesis testing procedure can be increased by increasing its Type I error. For the four items with low discrimination rates, the Mantel and SMD procedures had much greater Type

I error rates (recall Table 4) coupled with much lower power on average.

Mazzeo and Chang (1994) plotted the histograms of the 1000 SMD statistics indices calculated from the 1000 replications for the 11 DIF conditions listed in Table 6. They found that the distributional histograms shape reasonably normal looking. The increase in detecting DIF against the focal group when the a-parameter is large (false gains) was caused by the large shift of the testing statistic's distribution to the right of 0. In contrast, the large loss in power to detecting DIF against the focal group when the a-parameter small was due to the large shift the distribution to the left of 0. Such shifts which are closely related to the magnitude of the a-parameter sometimes cause Mantel and SMD to lose all power, e.g., when a=0.25 (see Table 6). It should be mentioned again that the Mantel and SMD are two very similar procedures. The main difference is that the SMD generates approximate N(0,1) statistics, while the Mantel generates approximate $\chi^2(1)$ statistics. The above explanation clearly illustrates that the direction of the DIF detected by these two procedures is controlled by a_j magnitudes.

5. Discussion

The paper discussed a modified version of the SIBTEST procedure for use with ordinal polytomous items and presented two simulation studies to illustrate the efficiency of the modified procedure. It is very interesting to point out that only two modifications to the original SIBTEST procedure were needed to make it applicable to polytomously scored items, as well as with tests consisting of both dichotomous and polytomous items. The simulation results suggest that these modifications made to the SIBTEST procedure have been essentially successful. It should be re-iterated that the unique correspondence between IRF and ICRFs (Chang & Mazzeo, 1994) for the commonly used polytomous IRT models is a useful guide for extending IRT dichotomous item DIF procedures to the polytomous item context. The natural generalization of an IRT/parametric-based definition of null DIF for an ordinally scored polytomous item is to require that item parameters are invariant across the two groups under study. But the Chang and Mazzeo result shows that an item score based definition is equivalent for commonly used polytomous models, thus justifying the definition used by the modified SIBTEST.

A hypothesis testing procedure is judged primarily by its power curve, that is, in the DIF context, by its ability to avoid false flagging (fixed low Type I error rate) of non-DIF items while detecting DIF items as much power (low Type II error) as possible. The simulation results here clearly demonstrate the superiority of the SIBTEST method of controlling impact-induced Type I error. Indeed the second study clearly indicates that the Mantel and SMD procedures may yield unacceptable results from both the Type I error and Type II error perspectives for DIF analyses of polytomous data even when the studied item is included in the matching test. The theoretical reason for this kind of error inflation can be obtained by a generalization of the well documented dichotomous work by Zwick, 1990; Meredith and Millsap, 1992; and Fischer, 1993. It should be mentioned that, according to a study by Allen and Donoghue (1994), for larger but realistic levels of variation in studied polytomous items discriminations, the impact induced Type I error inflation of the Mantel procedure is even more serious for shorter length tests ($\leq$ 20 items).

In contrast, SIBTEST exhibited only small changes in Type I error rates as a function of studied-item discrimination parameters. Hence, it appears that SIBTEST is much more robust to violation of Rasch model conditions. The current study shows that a small but consistent Type I error inflation does occur even with SIBTEST. Thus, a modification of the SIBTEST regression correction should be developed to better control the impact-related Type I error inflation. Such a research project has been begun.

There are two limitations in the current simulation design. First, the sample sizes of both the reference and focal groups are always the same. Second, in order to measure Type I error and power with total statistical accuracy, it was important to exclude DIF items from the matching test. In future studies, unequal sample sizes, and a robustness study of matching subtest contaminated with DIF items, and as well as detection of a group of DIF items (DTF, see Shealy & Stout, 1993) will be considered.

References

Allen, N. & Donoghue, J. (April, 1994). *DIF analysis based on complex samples of dichotomous and polytomous items.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, Louisiana.

Chang, H. (1994) A note on the monotonicity of the IRFs for polytomous IRT models. Unpublished Manuscript. Educational Testing Service, Princeton, NJ.

Chang, H., & Mazzeo, J. (1994). The unique correspondence of the item response function and item category response functions in polytomously scored item response models. *Psychometrika, 59,* 391-404.

Donoghue, J., Holland, P., & Thayer, D. (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.* Hillsdale, NJ: Lawrence Erlbaum Associates.

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., & Potenza, M.T. (1994). *Equity assessment for polytomously scored items: a taxonomy of procedures for assessing differential item functioning.* ETS Research Report (RR-94-49). Princeton, NJ: Educational Testing Service.

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

Fischer, G. (1993). Notes on the Mantel-Haenszel procedure and another chi-squared test for the assessment of DIF. *Sonderdruck Methodika 7,* 88-100.

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

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

Johson, E.G. & Carlson, J.E. (1994) *The NAEP 1992 technical report.* D.C.: National Center for Education Statistics.

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

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

Mazzeo, J., & Chang, H. (April, 1994). *Detecting DIF for polytomously scored items: progress made on the extension of Sealy-Stout's SIBTEST procedure.* Paper presented at 1994 AERA Annual Meeting, New Orleans, Louisiana.

Meredith, W. & Millsap, R.E. (1992). On tht misuse of manifest variables in the detection of measurement bias. *Psychometrika, 57,* 289-211.

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

Mullis, I., Dossey, J., Owen, E., & Phillips, G. (1993). *NAEP 1992 mathematics report card for the nation and the states.* Washington, DC: National Center for Education Statistics.

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

Roussos, L. & Stout, W. (1994). Simulation studies of effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type I error performance. *Journal of Educational Measurement*, under revision.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph No 17, 34* (4, Pt.2).

Shealy, R. & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item-bias/DIF. *Psychometrika, 58*, 159-194.

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

Thissen, D., Steinberg, L. & Wainer, H. (1993). An item response theory model for test bias and differential test functioning. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Welch, C., & Hoover, H.D. (1993). Procedures for extending item bias detection techniques to polytomously scored items. *Applied Measurement in Education, 6*, 1-20.

Zwick, R. (1990). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide?, *Journal of Educational Statistics, 15*, 185-197.

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

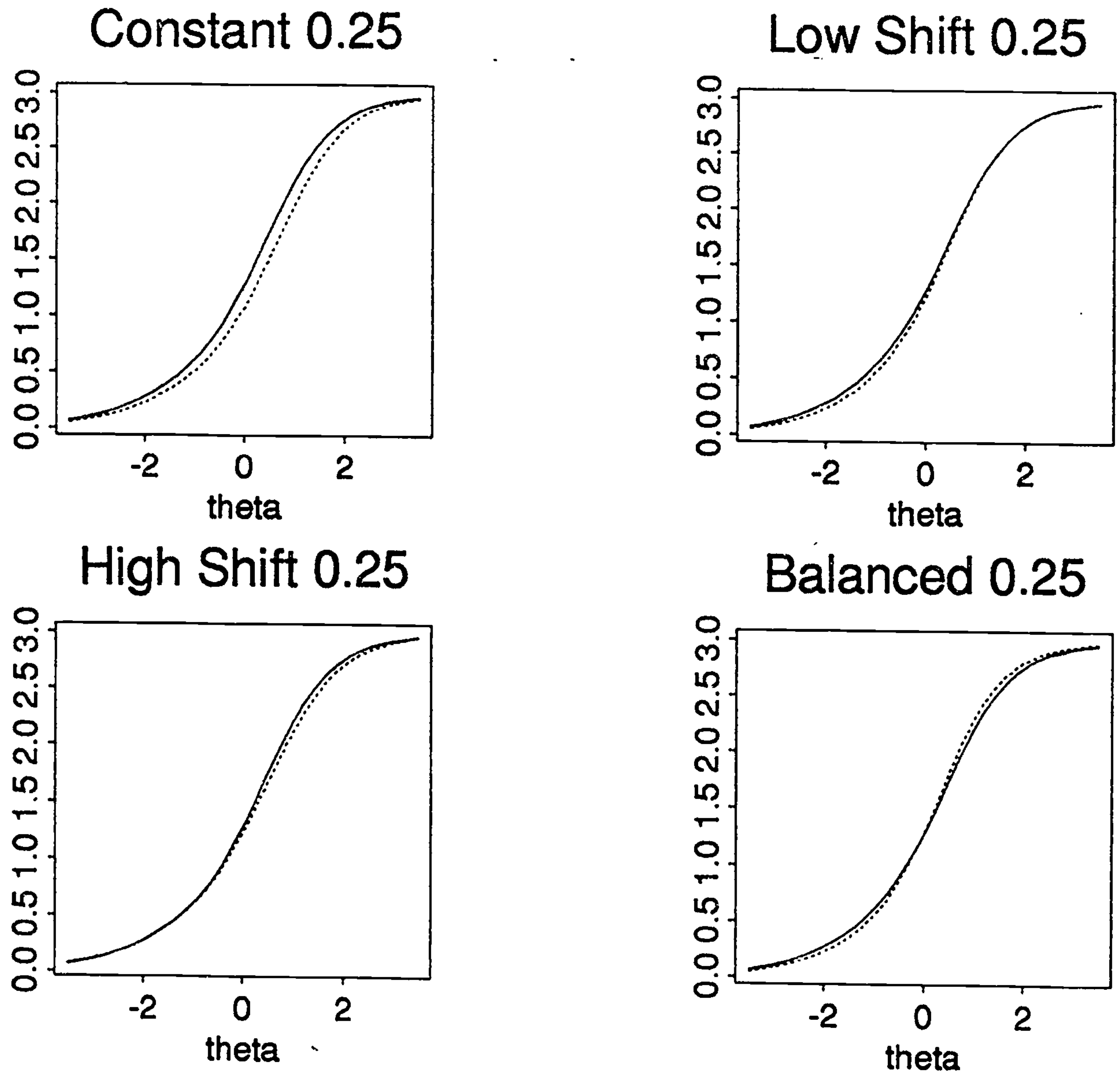

Figure 1: IRFs of Studied Item 1 for the Reference and Focal Groups for the 4 DIF Conditions (C=0.25).— Ref. IRF, ...Foc. IRF.