GRE®
RESEARCH

Describing and Categorizing DIF in Polytomous Items

Rebecca Zwick
Dorothy T. Thayer
and
John Mazzeo

May 1997

GRE Board Professional Report No. 93-10P
ETS Research Report 97-05

Educational Testing Service, Princeton, New Jersey

Describing and Categorizing DIF

in Polytomous Items

Rebecca Zwick
Dorothy T. Thayer
and
John Mazzeo

GRE Board Report No. 93-10P

May 1997

This report presents the findings of a research project funded by and carried out under the auspices of Educational Testing Service, the College Board Advanced Placement Program, and the Graduate Record Examinations Board.

Educational Testing Service, Princeton, NJ 08541

****

Researchers are encouraged to express freely their professional judgment. Therefore, points of view or opinions stated in Graduate Record Examinations Board Reports do not necessarily represent official Graduate Record Examinations Board position or policy.

****

The Graduate Record Examinations Board and Educational Testing Service are dedicated to the principle of equal opportunity, and their programs, services, and employment policies are guided by that principle.

Acknowledgments

We are grateful to the sponsors of this project: the Graduate Record Examinations Board, the Educational Testing Service Research Division, and the College Board Advanced Placement Program. John Mazzeo, who participated in the design and conduct of the simulation study, received additional support from the ETS research allocation funds of the Large-scale Assessment Group. We also thank the Graduate Record Examinations, Advanced Placement, and Praxis programs for generously providing us with the data used in the examples. We appreciate the contributions of Elana Broch, Marisa Farnum, Valerie Folk, Behroz Maneckshana, and Rick Morgan, who provided files and test items and responded to our many questions; Hua-Hua Chang and William Stout, who provided the SIBTEST code and consulted on its use; Jo-Lin Liang, who assisted with analyses; and Edward Kulick, who provided NAEP analysis results. Finally, we thank John Donoghue, Neil Dorans, and Nancy Petersen for their helpful reviews of the paper.

Abstract

The purpose of this project was to evaluate statistical procedures for assessing differential item functioning (DIF) in polytomous items (items with more than two score categories). Three descriptive statistics--the Standardized Mean Difference, or SMD (Dorans & Schmitt, 1991), and two procedures based on SIBTEST (Shealy & Stout, 1993) were considered, along with five inferential procedures--two based on SMD, two based on SIBTEST, and the Mantel (1963) method. The DIF procedures were evaluated through applications to simulated data, as well as data from ETS tests.

The simulation included conditions in which the two groups of examinees had the same ability distribution and conditions in which the group means differed by one standard deviation. When the two groups had the same distribution, the descriptive index that performed best was the SMD. When the two groups had different distributions, a modified form of the SIBTEST DIF effect size measure tended to perform best. The five inferential procedures performed almost indistinguishably when the two groups had identical distributions. When the two groups had different distributions and the studied item was highly discriminating, the SIBTEST procedures showed much better Type I error control than did the SMD and Mantel methods, particularly in short tests. The power ranking of the five procedures was inconsistent; it depended on the direction of DIF and other factors.

Routine application of these polytomous DIF methods at ETS seems feasible in cases where a reliable test is available for matching examinees. For the Mantel and SMD methods, Type I error control may be a concern under certain conditions. In the case of SIBTEST, the current version cannot easily accommodate matching tests that do not use number-right scoring. Additional research in these areas is likely to be useful.

1. Overview

Items with more than two score categories are referred to as polytomous items. Although the response categories for a polytomous item may be unordered, this study addressed only the case of items that are scored on an ordinal scale, such as essay items. At present, ETS has no official policy for screening polytomous items for DIF. Several promising statistical tests for DIF in polytomous items have been investigated (see reviews in Potenza & Dorans, 1995; Zwick & Thayer, 1994, in press; Welch & Miller, 1995), but little work has been conducted on the development of descriptive measures of DIF for these items. This project involved both the evaluation of inferential procedures and the development of DIF indexes and their standard errors. All methods selected for inclusion were judged to be relatively simple to apply in large-volume testing programs; none require the estimation of item parameters. Efforts were also made toward the establishment of a system for categorizing the severity of DIF which is analogous to the ETS system for characterizing dichotomous items as having no DIF ("A"), moderate DIF ("B"), or large DIF ("C"). It is hoped that this work will facilitate the implementation of DIF screening procedures for essay items and other free-response tasks that are scored on an ordinal scale.

The study focused on essentially unidimensional exams that include both multiple-choice and free-response components. In this type of test, DIF analysis of free-response items can be accomplished by matching examinees on the score on the multiple-choice section or the combined score for both sections. This type of analysis approach is also possible if there exists a suitable matching variable entirely external to the test in question.

Section 2 of this report gives a psychometric perspective on the issue of matching examinees in DIF analysis. Section 3 describes the DIF assessment procedures that were examined in this project. Section 4 gives information about the design and results of a simulation study that investigated the properties of these statistical methods. Section 5 provides examples of applications of polytomous DIF methods to three sets of ETS data: the free-response items from the 1993 College Board Advanced Placement (AP®) Physics B exam, the new Graduate Record Examinations (GRE®) writing measure, and essays from the Praxis I writing assessment. Section 6 addresses the development of DIF categories for polytomous items. Finally, Section 7 provides a summary and discussion. With a few exceptions, formulas are not presented in this report; these can be found in the cited references.

2. Errors in matching examinees in DIF assessment

All the DIF methods evaluated in this project involve matching examinees from two groups, such as men and women, on a test score (or adjusted test score) and then comparing the item performance of the matched members of the two groups. In general, the validity of the DIF assessment results depends on the accurate matching of the examinees; large or systematic error in the matching variable can lead to incorrect conclusions.[1] The possible effects of errors in matching can be considered from a classical test theory perspective (see also Shealy & Stout, 1993; Zwick, 1990). Let S be an observed test score, let T be the corresponding true score, and let E represent measurement error. (Note that T is a latent ability variable expressed in the test score metric; except for scale of measurement, it is the same as the usual "θ" of item response theory.) Under the classical test theory model (see Lord & Novick, 1968, chapter 3),

$$S = T + E \quad \text{and} \quad E(S) = T.$$

The linear least squares regression function for the regression of T on S is

$$E(T|S=s) = \rho s + (1 - \rho)\ \mu\ , \qquad (1)$$

where ρ = Var(T)/ Var(S) is the reliability of the test, and $\mu = E(T) = E(S)$. (Further assumptions are needed in order to claim that the *true* regression is linear; see Lord & Novick, 1968, p. 65; pp. 503-505.)

Equation 1 is sometimes referred to as the "Kelley correction" because T. L. Kelley (1923) proposed an equivalent formula as a means of adjusting an observed test score for measurement error. If the test reliability, ρ, is equal to one, then $E(T|S=s)$ is equal to s (the expected true score is equal to the observed score), whereas if ρ is equal to zero, then $E(T|S=s)$ is equal to μ (the expected true score is equal to the mean of the true scores, which is equal to the mean of the observed scores). In general, the lower the test reliability, the more the expected true score regresses toward the mean.

Now consider the case of DIF analysis, in which we compare the group of primary interest, or *focal group*, to a *reference group*. Suppose that the test

[1] For a different perspective on this issue, see Potenza and Dorans (1995).

reliabilities for the two groups are equal (i.e., $\rho_F = \rho_R$), but that the mean test score for the focal population is less than the mean for the reference population ($\mu_F < \mu_R$). The result is that, for a particular observed score s, the expected true score for the focal group is less than the expected true score for the reference group ($E_F(T \mid S = s) < E_R(T \mid S = s)$). For example, suppose that $\rho_F = \rho_R = .85$, $\mu_F = 40$, and $\mu_R = 60$. Now consider an observed score of S = 50. Whereas $E_F(T \mid S = 50) = .85(50) + .15(40) = 48$, we find that $E_R(T \mid S = 50) = .85(50) + .15(60) = 52$. It is clear from equation 1 that, under the classical test theory model, this matching problem decreases as $|\mu_R - \mu_F|$ decreases and as ρ increases.

DIF procedures are affected in different ways by the errors in the matching variable. In particular, the SIBTEST procedure (Shealy & Stout, 1993), along with its extension to polytomous items (Chang, Mazzeo, & Roussos, in press) can be contrasted with the Mantel-Haenszel (1959) procedure (Holland & Thayer, 1988) and its polytomous-item extension (Mantel, 1963; see Zwick, Donoghue, & Grima, 1993a, 1993b).

The SIBTEST procedures include a correction for measurement error: The matching variable is computed by applying a Kelley-type adjustment to the number-right score on the matching test, excluding the studied item. Further details are provided in Section 3.1.2.

Although the Mantel-Haenszel (MH) and Mantel procedures are sometimes said to "correct for" measurement error by including the studied item in the matching variable, this is, in fact, a misleading description. These methods include *no explicit correction*, but, fortuitously, the measurement error problem can be shown to vanish in certain cases. Specifically, if examinees are matched on number-right score S, *including the studied item,* and if S is a sufficient statistic for T, then measurement error in S does not affect DIF results. (For a discussion of sufficient statistics, see Lord & Novick, 1968, chapter 18.) In essence, sufficiency of S for T (which is equivalent to sufficiency of S for θ) means that S contains all the information about T (or θ) in the item response data. Number-right score is a sufficient statistic for ability when all items in the test follow the Rasch model or partial credit model (Masters, 1982; see equation 3 in Section 4.3). In order for sufficiency to hold, it is also necessary to assume that all items, with the possible exception of the item under study, are free of DIF and that item responses are independent, given θ. Sufficiency does not hold (even in the Rasch and partial credit models) if the studied item is not included in the score used for matching.

Theoretical discussions of the special relationship of the MH to the Rasch model are given in Fischer (1993), Holland and Thayer (1988), Lewis (1993), Meredith and Millsap (1992), Millsap and Meredith (1992), Zwick (1990), and Zwick, Donoghue, and Grima (1993a, 1993b). In a particularly useful paper, Lewis (1993) shows that the sufficiency of S in the Rasch case eliminates the problem of measurement error for a reason that is unrelated to test reliability. Even if S is unreliable, matching on S in this special case renders the studied item score independent of θ so that a group difference in the distribution of θ (which was shown to result in a measurement error problem in the example of Section 2) becomes irrelevant: In the Rasch case, the probability of a correct item response, given a particular observed test score, does not depend on θ.

Measurement error affects other DIF procedures that match on observed test scores in essentially the same way as it affects the MH and Mantel methods. This category of procedures includes the standardization approach (Dorans & Kulick, 1986; Dorans & Schmitt, 1991), which is discussed in Section 3.

From a practical perspective, a key question is: How robust are methods that match on observed score to departures from the Rasch model? Simulation results have been mixed. Some simulation studies that have shown considerable robustness of observed-score methods are Donoghue, Holland, and Thayer (1993) and Shealy and Stout (1993) in the dichotomous case and Zwick, Donoghue, and Grima (1993a, 1993b) in the polytomous case. Other researchers have found very large Type I error rates for the observed-score methods under certain conditions, including Shealy (1989), Roussos and Stout (1993), and Uttaro and Millsap (1994) in the case of the MH and Allen and Donoghue (in press) and Chang, Mazzeo, and Roussos (in press) in the case of the Mantel or SMD procedures. Further discussion of the robustness issue appears in Section 4.3.1.

3. Polytomous DIF procedures

Three descriptive procedures and five inferential procedures were examined in this study.

3.1. Descriptive procedures

The three descriptive measures of DIF that were studied in the simulation were the standardized mean difference (SMD; see Dorans & Schmitt, 1991 and Zwick & Thayer, 1994, in press) and two alternative effect size measures based on the extension of the Shealy-Stout (1993) SIBTEST procedure to polytomous items (Chang, Mazzeo, & Roussos, in press). The test statistic used in the SIBTEST inferential procedure is obtained by dividing a DIF effect measure by its standard error. The SIBTEST effect size measures are simply the numerators

of the SIBTEST statistics described in Section 3.2. The three indexes, all of which express DIF in the same metric as the item score, are described in detail below.

3.1.1. SMD

Dorans and Schmitt (1991) proposed a descriptive index that compares the item means of two groups, after adjusting for differences in the distribution of members of the two groups across the values of the matching variable. In the case of SMD, the matching variable is defined as the sum of the score on the matching items and the score on the studied item. The statistic is an extension of STD P-DIF, a standardized difference in proportions correct developed by Dorans and Kulick (1986) for dichotomous items. SMD reduces to STD P-DIF in the dichotomous case.

3.1.2. Standard SIBTEST DIF measure

As noted in Section 2, the SIBTEST procedure of Shealy and Stout (1993) and its extension to polytomous items (Chang, Mazzeo & Roussos, in press) include an explicit correction for measurement error in the matching variable. The correction is obtained by assuming the model in equation 1 for the regression of true score on observed score (which, in this case, is the number-right score on the matching items, excluding the studied item). The slope of this regression is the reliability of the test (excluding the studied item). Shealy and Stout (1993, p. 192, part 5) use the following estimator of the reliability (computed separately for the reference and focal groups):

$$Reliability = \left(\frac{n}{n-1}\right)\left\{1 - \frac{\sum_{i=1}^{n} p_i^*\left(1-p_i^*\right)}{Total\ test\ variance}\right\} \qquad (2)$$

In equation 2, n is the number of items and $p_i^* = \left(p_i - c\right)/(1-c)$, where p_i is the observed proportion correct for item i and c is a user-specified guessing parameter for the test. This is identical to Cronbach's alpha, except that in computing the item variances in the numerator, a "guessing correction" is applied. The value of c in the analyses of our simulation data was .14.[2]

2 The c value is intended to be the average probability of obtaining a correct response on the matching items by guessing. Our estimate of .14 was obtained by averaging the estimated guessing parameters for the dichotomous matching items, which had been obtained in a previous

3.1.3. Modified SIBTEST DIF measure

This modified version of the Standard SIBTEST DIF measure was obtained by omitting the guessing correction (i.e., setting c to zero in equation 2.)[3]

3.2. Inferential procedures

Two of the five inferential procedures that were investigated are based on the SMD. These methods make use of two standard error formulas derived by Zwick (1992; Zwick & Thayer, 1994, in press). One standard error is based on a hypergeometric model, the other on a multinomial model. Earlier simulations showed that the hypergeometric standard error performed better than the multinomial standard error.[4] The Mantel procedure and two SIBTEST procedures were also investigated. Each of the five methods, which are described below, yields a test statistic that is approximately standard normal under the null hypothesis of no DIF.

---

simulation study (see Section 4.2). This seemed preferable to using the generating value of the guessing parameter (.15), which would not ordinarily be known to users.

3 A simulation showed that applying the guessing correction impaired estimation of the true reliability (approximated by computing the true and observed score variances from a simulation based on 5,000 observations). That is, in the majority of cases, the ordinary Cronbach's alpha coefficient was closer to the true reliability than was the guessing-corrected version. However, Stout (personal communication, October 1994) advised against deleting the correction, noting that it had been shown to improve the Type I error performance of SIBTEST's inferential procedure. We therefore included both the standard guessing-corrected SIBTEST procedure and this modified version. A new version of SIBTEST (Douglas, Stout, & DiBello, 1995) uses an algorithm that does not involve test reliability.

4 The hypergeometric standard error was derived by assuming that, under the null hypothesis of no conditional association between item score and group membership, the vector of item score frequencies for one group (say, the focal group) has a multivariate hypergeometric distribution within each level of the matching variable. This assumption is the same as that implicitly used by Mantel (1963); see Section 3.2.3. The multinomial standard error was derived by assuming that, within each level of the matching variable, the vector of item score frequencies for each group has a multinomial distribution. The multinomial parameters for the reference and focal groups were not assumed to be identical. Details appear in Zwick and Thayer (1994, in press). As a spin-off of this research, an investigation was conducted of standard error formulas for the dichotomous case, when SMD is equal to STD P-DIF (Dorans & Kulick, 1986). Findings to date suggest that the hypergeometric standard error formula developed for SMD performs better than the standard error formula for STD P-DIF which is now used by ETS testing programs (see Zwick & Thayer, 1995).

3.2.1. SMD-H

The SMD-H test is obtained by dividing the SMD statistic of Section 3.1.1 by a standard error derived under a hypergeometric model.

3.2.2. SMD-M

The SMD-M test is obtained by dividing the SMD statistic of Section 3.1.1 by a standard error derived under the multinomial model. This standard error was shown to be very similar to the SIBTEST standard error of Shealy and Stout (1993, p. 169, equation 19).

3.2.3. Mantel's procedure

Zwick, Donoghue, and Grima (1993a, 1993b) described a DIF analysis approach for polytomous items which is based on a statistical test developed by Mantel (1963). In the case of dichotomous items, Mantel's statistic reduces to the Mantel-Haenszel (1959) test (without the continuity correction), which is the basis for the DIF procedure of Holland and Thayer (1988). As in the case of SMD, the matching variable is defined as the sum of the scores on the matching items and the score on the studied item. The procedure has been applied for DIF purposes by several researchers (see Potenza & Dorans, 1995; Zwick & Thayer, 1994, in press; Welch & Miller, 1995).

3.2.4. Standard SIBTEST

This test is obtained by dividing the Standard SIBTEST DIF measure of Section 3.1.2 by its standard error, as described by Shealy and Stout (1993).

3.2.5. Modified SIBTEST

This test is obtained by dividing the Modified SIBTEST DIF measure of Section 3.1.3 by the Shealy-Stout (1993) standard error. (The standard error is not affected by inclusion of the guessing correction.)

3.3. Unified implementation decisions for DIF procedures

We implemented the various DIF methods as similarly as possible so that the comparisons across procedures would not be affected by extraneous factors (e.g., treatment of missing data). The following decisions were part of this unified implementation strategy:

(a) To keep the signs of all DIF statistics consistent, the SIBTEST effect measures and the associated Z statistics were "reflected" so that a negative value indicated DIF against the focal group.

(b) To achieve the "standardization" that is a feature of the SMD and SIBTEST procedures, *focal group weighting* was used. This means that the between-group difference in mean item score for examinees at, say, level *k* of the matching variable was weighted by the proportion of *focal group members* who were at level *k*. This is consistent with the approach that is used in ETS applications of standardization DIF procedures. In regard to the *inferential* application of SIBTEST, Shealy and Stout (1993, p. 176) reported slightly better Type I error performance when weights based on the total group were used. They reported that power results were very similar for the two types of weights.
(c) For all procedures, missing data were treated as detailed in the appendix of the Shealy-Stout (1993, pp. 190-193) description of the SIBTEST procedure.[5] The Shealy-Stout missing data rules provide for the elimination of data at the extremes of the matching variable and at levels of the matching variable for which the number of observations is less than a user-specified minimum. (This minimum was set to 2 in our study.) In our simulation, these rules resulted in the elimination of 2% to 4% of the cases when the two groups of examinees had the same ability distribution and 4% to 5% of the cases when the groups had different distributions.

4. Simulation study

A major component of this project was a large-scale simulation study that evaluated the behavior of the three descriptive and five inferential procedures described in Section 3. The study was more comprehensive than previous simulations of polytomous DIF measures, such as that of Zwick, Donoghue, and Grima (1993a, 1993b), in that it (a) included a larger number of statistical procedures (both descriptive and inferential), (b) examined the effect of jointly varying item discrimination and difficulty, and (c) considered both DIF against the focal group and DIF against the reference group.

4.1. Overall design of data generation and analyses

The factors considered in the simulation were focal group population (2 levels), matching test length (2 levels), and three properties of the studied item: discrimination (a; 3 levels), reference group difficulty (b_R; 2 levels), and difference between reference and focal group difficulty (d; 3 levels). Crossing all these factors yields 72 cells.

---

[5] An exception was that rule 4B was not implemented. The rule states that the data at a particular level of the matching variable should be discarded from the analysis if the item score variance is zero in either the reference or the focal group. Stout (personal communication, October 1994) noted that rule 4B is no longer recommended and is not part of the SIBTEST code.

In all cases, the reference group was drawn from a standard normal (N(0, 1)) distribution. Two focal populations were included: N(0, 1) and N(-1, 1). As described in Section 2, all methods were expected to perform well when the reference and focal populations were the same (the N(0, 1) condition). The N(-1, 1) condition, which represents a situation in which measurement error is a potential problem, was expected to reveal differences in the effectiveness of the procedures.

Each examinee record consisted of an ability drawn from one of these distributions, responses to 50 dichotomous matching items, and responses to 18 polytomous studied items. (Models for generating the item responses are described below.) In half of the DIF analyses, all 50 dichotomous items were used for matching; in the remainder of the analyses, the matching test consisted of a 20-item subset of the 50 items. Each studied item was considered in turn; that is, each DIF analysis was based on a set of either 20 or 50 matching items, along with one studied item. As explained in Section 2, any distortions in the functioning of the statistical procedures were expected to be less severe with the 50-item matching test (when reliability is higher) than with the 20-item matching test. Each DIF analysis was repeated 500 times, with samples of 500 reference group and 500 focal group examinees.

The design is summarized in the following table:

	20-item matching test	50-item matching test
Focal population N(0, 1)	18 studied items*	18 studied items
Focal population N(-1, 1)	18 studied items	18 studied items

*formed by crossing discrimination (3 levels), reference group difficulty (2 levels), and difference between reference and focal group difficulty (3 levels).

4.2. Matching items

The responses to the dichotomous matching items were generated using the three-parameter logistic model. The matching items were subsets of the 75 items used in the simulation conducted by Zwick, Thayer, and Wingersky (1994). (The selection of item parameter values for this earlier study was based on analyses of actual test data.) The discrimination (*a*) parameters for the matching items were either .74 or 1, with an average of .86, the difficulty (*b*) parameters ranged from -1.95 to 1.95, with a mean of 0, and the guessing parameters were equal to .15. The 20- and 50-item subsets used in the present study were chosen to have the

same average discrimination and difficulty as the original set of 75 items. The matching items contained no DIF.

4.3. Studied items

Eighteen distinct studied items were generated using the generalized partial credit model (Muraki, 1992). Assuming that there are $M + 1$ score categories (0, 1, ... M), the probability of receiving a score x on item i for an examinee in group g with proficiency θ is given by

$$P_{xig}(\theta) = \frac{\exp \sum_{m=0}^{x} 1.7a_i(\theta - b_{ig} - \delta_{mi})}{\sum_{p=0}^{M} \exp \sum_{m=0}^{p} 1.7a_i(\theta - b_{ig} - \delta_{mi})}, \quad x = 0, 1, \ldots M. \qquad (3)$$

In equation 3, which is known as the *item category response function*,

θ is the examinee ability,

a_i is the item discrimination,

b_{ig} is the overall item difficulty in group g, with $b_{iF} = b_{iR} - d_i$ (see section 4.3.3), and

δ_{mi} is the difficulty of making the transition from category $m-1$ to category m. For all studied items in the simulation, the following δ_{mi} values were used: $\delta_{1i} = -.75$, $\delta_{2i} = 0$, $\delta_{3i} = .75$. The term $\sum_{m=0}^{0} a_i(\theta - b_{ig} - \delta_{mi})$ is assumed to be equal to zero by convention. M was equal to 3; that is, the items were "scored" on a scale ranging from 0 to 3.

The three factors (a, b_R, and d) that were crossed to produce the 18 studied items are described in Sections 4.3.1 through 4.3.3. Table 1 lists the parameter values for each item. The table also includes the true DIF values for each item, obtained by computing the between-group difference in the expected item score, given θ, and then integrating that conditional difference over the focal group ability distribution (say, $g_F(\theta)$). That is, the true DIF for item i is defined as follows:

$$True\ DIF = \int \left(\sum_{x=0}^{M} xP_{xiF}(\theta) - \sum_{x=0}^{M} xP_{xiR}(\theta) \right) g_F(\theta) d\theta, \qquad (4)$$

where x indexes the item response categories and $P_{xig}(\theta)$ is given by equation 3. Ideally, the DIF effect size measures in Section 3 would be equal to the true DIF values for each item.

4.3.1. Item discrimination (a)

Three levels of item discrimination were included: .47, .86, and 1.57. Previous theoretical and simulation work has demonstrated the relevance of this factor to DIF conclusions. When all items in a test follow the Rasch or partial credit model, total test score is a sufficient statistic for θ; DIF procedures that match on number-right score work best under these circumstances (see Section 2). The performance of these methods can be degraded when the item response model departs from the Rasch case. In particular, Chang, Mazzeo, and Roussos (in press) and Mazzeo and Chang (1994) showed that when the studied item discrimination parameter is substantially higher than the item discrimination parameters of the matching items, the Mantel method and other observed-score procedures can produce distorted results, including unacceptably high Type I error rates. These results parallel those of Roussos and Stout (1993) and Uttaro and Millsap (1994) for the dichotomous case.

One of the three studied item a values in the current simulation was set equal to the average a value for the matching items (.86). The other two a values, .47 and 1.57, depart from this average by equal amounts in the natural log metric. Type I error rates for the observed-score methods were expected to be higher for a = .47 and a = 1.57 than for a = .86. That is, it was assumed that the most "Rasch-like" condition in the study was the one in which the studied item a was equal to .86.

4.3.2. Item difficulty in the reference group (b_R)

Previous studies of DIF procedures for polytomous items (e.g. Mazzeo & Chang, 1994 and Zwick, Donoghue, & Grima, 1993a, 1993b) did not incorporate a systematic investigation of the effects of jointly varying the discrimination and difficulty parameters of the studied items. The present study crossed two levels of b_R (-.5 and .5) with the three levels of a.

4.3.3. Difference in reference and focal group difficulty (d)

The d parameter represents the difference between the reference and focal group difficulty and is therefore directly related to the amount of DIF. Three levels of d were included: -.25 (DIF against the focal group), 0 (no DIF), and .25 (DIF against the reference group). The effective amount of DIF in an item is

also influenced by the a parameter. In general, given the focal group distribution and the values of d and b_R, the size of the true DIF of equation 4 increases with a (see Table 1). Theoretical results and findings from previous simulations lead us to believe that $|d| = .25$ is realistic. It is the larger of two d values included in the simulations of Zwick, Donoghue, and Grima (1993a, 1993b) and Chang, Mazzeo, and Roussos (in press).

4.4. Results for descriptive procedures

Section 4.4.1 presents findings about the factors that affected the size of the descriptive measures of DIF. Section 4.4.2 evaluates the DIF measures in terms of their departures from the true DIF values for each item.

4.4.1 Factors affecting the magnitude of DIF measures

Analyses of variance (ANOVAs) were conducted to determine which factors had the most impact on the effect size measures. DIF effect size was the dependent variable and a procedure indicator (P) was included as a design variable to identify the three measures. The other design variables were the factors defined above: focal population (F), matching test length (L), item discrimination (A), item difficulty (B), difference between reference and focal group difficulties (D) and all interactions up to and including three-way interactions. Higher-order interactions were included in the residual term. One ANOVA included only the studied items with no DIF. (D was, of course, excluded as a design variable in this analysis.) A second ANOVA included all studied items. Because the number of observations was very large, even tiny effects were statistically significant. Therefore, for interpretation purposes, emphasis was placed on the factors that explained at least 1% of the variance in the effect size measures.

For no-DIF items, it is undesirable for any of the design factors to affect the observed DIF measures: All observed statistics should ideally be equal to zero. In the ANOVA for these items, the largest effects were P, F, A x P, F x P, and A x F x P, each explaining 1% to 2% of the variance. Because of the interactions involving Procedure, ANOVAs that examined each of the three DIF measures separately proved to be more useful for interpretation: For SMD, four terms explained at least 1% of the variance: A, F, and A x F (each explaining 6% to 7%) and B x F (1%). For Standard SIBTEST, 2% of the variance was attributable to F, and for Modified SIBTEST, no term explained even 1% of the variance.

In the ANOVAs that included all descriptive procedures, items, and conditions, the only terms that explained at least 1% of the variance were D (82%) and A x D (1%). Because interactions with Procedure were trivially

small, nearly identical results were obtained by examining each of the three effect size measures separately. As noted in Section 4.3.3, the effective amount of DIF in an item is influenced by the a parameter; therefore, the obtained contributions of both D and A x D are consistent with theory.

4.4.2. Differences between DIF measures and true DIF values

To better illustrate the behavior of the descriptive measures, residuals were computed by subtracting the true DIF values from Table 1 from the observed values. The true DIF values, computed from equation 4, are the estimands--the quantities estimated by the observed statistics. For no-DIF items, the true DIF values are zero, and the residuals are equal to the DIF measures themselves. Ideally, all the residuals for all items would be equal to zero.

Table 2 shows the average residuals for the three DIF effect measures for each value of a and each focal group condition. Each mean, which is averaged over b_R and matching test length, is based on 2,000 observations. The standard errors are approximately .001 for most of the table entries. For the N(-1, 1) condition, $d \pm .25$, some standard errors are slightly larger, but none exceed .002. As detailed below, the performance of the three DIF measures was affected in complex ways by the design factors and was therefore very difficult to characterize.

First, consider the no-DIF items, for which the average DIF measures (average residuals) are shown in the middle portion of the table. As in previous simulations, all measures functioned well when the focal group mean was equal to the reference group mean (the N(0, 1) focal group condition); there were no differences across methods. When the population means differed (the N(-1, 1) condition), the mean effect sizes departed from zero to varying degrees. For SMD, Standard SIBTEST, and Modified SIBTEST, respectively, the maximal departures (for cells defined as in Table 2) were -.059, -.020, and .005. SMD had its largest departure (-.059) at $a = 1.57$; it performed best at $a = .47$. This was contrary to the conjecture (Section 4.3.1) that the best performance of observed-score methods would occur when the studied item discrimination was the same as the average discrimination of the matching items (.86). For Standard SIBTEST, all six averages were negative, and departures from zero were always at least as large as those for Modified SIBTEST.

When DIF was present and the focal group distribution was N(0, 1), SMD showed departures from zero that were consistently smaller than those of the SIBTEST measures. In the N(-1, 1) focal group condition, Modified SIBTEST usually showed smaller departures than the other two statistics. SMD had three departures of at least .05 in magnitude, including a value of -.107 for $d = .25$, $a =$

1.57. SMD was also shown to be quite asymmetric in that, for a = .86 and a = 1.57, its departures from zero were much greater for d = .25 than for d = -.25.

Table 3 shows the average residuals for each value of b_R and d for each focal group condition. Each cell is averaged over a and matching test length and is based on 3,000 observations. The cell standard errors are less than .001. One readily apparent result is that, in general, departures from zero tended to be larger for b_R = .5 than for b_R = -.5. In the N(0, 1) focal group condition, there was again evidence that SMD outperformed its competitors, whereas the SIBTEST procedures tended to perform better than SMD in the N(-1, 1) focal group condition. SMD again showed the largest departure from zero--a cell mean of -.085 for d = .25, b_R = .5--and again behaved in an asymmetric way.

4.5. Results for inferential procedures

Section 4.5.1 gives findings about the factors that affected the rates with which the null hypothesis of no DIF was rejected for each of the inferential procedures. The rejection rates themselves are discussed in Section 4.5.2.

4.5.1. Factors affecting rejection rates

Logistic regression analyses were conducted to determine which factors had the most impact on the rates with which the null hypothesis of no DIF was rejected. In the first set of analyses, the outcome (reject or do not reject at α = .05) was the dependent variable and a procedure indicator (P) was included as a predictor to identify the five inferential procedures. The other predictors were the factors defined in the description of the ANOVAs in Section 4.4. One logistic regression included only the studied items with no DIF. (D was excluded as a predictor in this analysis.) A second logistic regression included all studied items.

Because of interactions between the procedure indicator and other predictors, logistic regressions that were conducted separately for each procedure proved easier to interpret. These analyses included all interactions up to and including three-way interactions. Because there is no satisfactory analogue to measures of explained variance in logistic regression, the impact of each effect was assessed using statistical significance as a criterion. Effects associated with a significance probability less than .005 were considered worthy of further examination and discussion. (This is in some sense a more inclusive criterion than that used for the ANOVAs in that all the effects with at least 1% of the variance were statistically significant at $p < .0001$.)

The top portion of Table 4 lists, for the no-DIF items, effects that were statistically significant ($p < .005$) for at least one of the inferential procedures. The lower portion of the table gives results for all items considered simultaneously. The body of the table indicates for which procedures each effect was significant. (A key to the effect abbreviations appears in the table footnote.)

It is undesirable for any of the predictors to affect the rejection rates for the no-DIF items. Table 4 shows, however, that for the Mantel, SMD-H, and SMD-M, the A (discrimination) x F (focal group distribution), A x B(reference group difficulty) x F, and A x F x L (test length) interactions met the statistical significance criterion, with more extreme results for A x F. For Standard and Modified SIBTEST, no term met the statistical significance criterion.

When all items were considered simultaneously, 13 effects were found to be statistically significant for at least one DIF procedure. Seven of these effects were significant for all five procedures: A, D, A x B, A x D, A x L, A x D x F, and B x D x L. As in the case of the DIF effect size measures, A (discrimination) and D (difference between reference and focal group difficulty) and their interactions with other factors were the most important influences.

4.5.2. Rejection rates

Table 5 gives results for each level of a in each focal group condition (averaged over b_R and matching test length). Each cell represents the rejection rate for 2,000 observations; cell standard errors range from .001 to .011. Table 6 provides a more detailed picture of results for the N(-1, 1) focal group condition. Rejection rates are shown for each level of a and matching test length (averaged over levels of b_R). Each cell represents the rejection rate for 1,000 observations; standard errors range from .001 to .016. The middle sections of Tables 5 and 6 (where $d = 0$) show the Type I error rates for the five inferential procedures.

Table 5 shows that all five procedures had Type I error rates close to their nominal value of .05 in the N(0, 1) focal group condition. However, in the N(-1, 1) focal group condition, differences among the procedures were evident. Type I error rates for all procedures were acceptable for $a = .47$, with the SIBTEST procedures showing trivially higher rates than the remaining three procedures. For $a = .86$, all five procedures had somewhat elevated rates (.12 to .13 for the SMD and Mantel procedures and .09 for the SIBTEST procedures), which is again contrary to the expectation that the observed-score methods would perform best at $a = .86$. The differences among the procedures were most dramatic for $a = 1.57$. Here, the Type I error rates for the Mantel and SMD procedures were much too high (.34 to .35), while the rates for the SIBTEST procedures were less

excessive (.11 to .13). Type I error rates for SMD-M were typically higher than those for SMD-H by .01. (Theoretical reasons for this are discussed in Zwick & Thayer, 1994, in press, who found SMD-M to have inflated Type I error rates in some conditions.)

In Table 6, the results for $d = 0$ provide some insight into the effect of test length on Type I error rate elevation for the SMD and Mantel methods. For a = .47, the Type I error rates for the Mantel and SMD methods never exceeded .06; Standard and Modified SIBTEST had slightly higher rates. For a = .86, rates for all methods were elevated, as noted earlier. For a matching test of 20 items, the error rates for Standard and Modified SIBTEST were lower than those for the Mantel and SMD methods; for a matching test of 50 items, error rates for the four methods were nearly identical. The largest differences across methods occurred for a = 1.57. Here, the Mantel and SMD methods had extremely high Type I error rates for a matching test length of 20 (.49 to .50); the SIBTEST procedures showed substantially lower rates (.12 to .16). For the 50-item matching test, the Mantel and SMD rates were reduced to .19 to .22--still higher than the rates of .09 to .10 for the SIBTEST procedures.

For items with DIF, Table 5 shows that all methods performed well in the N(0, 1) condition; their rejection rates were virtually indistinguishable. In the N(-1, 1) focal group condition, the power results were as complex as the Type I error findings. For a = .47, where all procedures had comparable Type I error rates (and therefore the comparison of rejection rates can legitimately be called a power comparison), the SIBTEST procedures tended to be less powerful than the competing methods for d = .25, but more powerful for d = -.25. Also, for d = .25, Modified SIBTEST was more powerful than SIBTEST, whereas for d = -.25, the reverse was true. SMD-H and SMD-M had similar rejection rates, typically differing by no more than .03. The Mantel method, Standard SIBTEST, and especially the two SMD procedures behaved asymmetrically in that their rejection rates for d = -.25 differed from their rejection rates for d = .25. For Standard SIBTEST, rates were higher for negative DIF (e.g., .68 for d = -.25, compared to .49 for d = .25, at a = .47). For SMD-H and SMD-M, the direction of the difference depended on a. Modified SIBTEST behaved in a more symmetric fashion than the other methods.

The power findings in Table 6 parallel those in Table 5 in most respects. As an example of the difficulty of ordering the procedures in terms of their power, consider the case in which the length of the matching test is 50 and a = .86. The Type I error rate was nearly the same across methods (.08 to .09) here. For d = -.25 (DIF against the focal group), the ordering of the procedures from most to least powerful was as follows: Mantel, SMD-M, SMD-H, Standard SIBTEST, Modified SIBTEST. For d = .25 (DIF against the reference group),

the ordering was Modified SIBTEST, Mantel, Standard SIBTEST, SMD-M, SMD-H. The asymmetric behavior of the procedures was again evident; Modified SIBTEST behaved most symmetrically.

4.6. Summary of results for descriptive and inferential procedures

In the N(0, 1) focal group condition, examination of residuals (obtained by subtracting true DIF from observed DIF) showed that SMD performed best as a descriptive statistic, whether or not DIF was present. In the N(-1, 1) focal group condition, however, Modified SIBTEST tended to have the smallest residuals. ANOVAs confirmed that, in the case of no-DIF items, Modified SIBTEST was the least affected by nuisance factors, followed by Standard SIBTEST and SMD in that order.

The five inferential procedures performed almost indistinguishably when the reference and focal groups had identical distributions. When the focal group had a N(-1, 1) distribution (i.e., the focal group mean was one standard deviation lower than the reference group mean) and the discrimination of the studied item was 1.57 (higher than the average discrimination of the matching items), the SIBTEST procedures showed much better Type I error control than the SMD and Mantel methods, which had extremely high error rates. Logistic regression showed that the Type I error rates for the SIBTEST procedures were less affected by nuisance factors than the error rates for the observed-score methods.

Test length had a substantial modifying influence on the Type I error inflation in the observed-score methods. For example, at $a = 1.57$, increasing the matching test length from 20 to 50 (and hence increasing the reliability of the matching variable from about .75 to about .85) reduced the Type I error rates for the SMD and Mantel methods from about .50 to about .20.

The power ranking of the five inferential procedures was inconsistent; it depended on the direction of DIF and other factors. In the N(-1, 1) focal group condition, the Mantel method, Standard SIBTEST, and especially the two SMD procedures behaved asymmetrically in that their rejection rates for $d = -.25$ differed from their rejection rates for $d = .25$. Modified SIBTEST behaved in a more symmetric way than its competitors.

5. Applications to ETS data

Several applications of polytomous DIF methods to ETS data sets were conducted. The College Board Advanced Placement (AP), Graduate Record Examinations (GRE), and Praxis programs generously allowed us to use their data to illustrate the various polytomous DIF methods. The limitations that

should be considered in interpreting the results of these experimental applications are discussed in Section 5.4.

The items that were analyzed were the free-response items from the AP Physics B exam, the new GRE writing measure, and the essays from the Praxis I: Computer-Based Test (CBT) writing assessment. In all three tests, the matching variable was based on the score on a multiple-choice exam. In the case of the GRE and Praxis tests, the multiple-choice items were given as computer-adaptive tests (CATs) and the matching variable was based on the CAT score, an expected true score derived through item response theory. (See Zwick, Thayer, & Wingersky, 1994, for a discussion of the use of this type of matching variable in DIF analysis.) In the version of the SIBTEST program available to us, it is assumed that number-right scoring can be applied to the matching test; therefore, it was not possible to apply SIBTEST to the GRE or Praxis I: CBT data. Only the Mantel and SMD procedures were used for these two data sets.

5.1. Analyses of 1993 AP Physics B data

Many of the College Board AP examinations, which are taken by high school students to determine college course placement, include both a multiple-choice and a free-response section. In the past, the free-response sections have not been subjected to DIF analysis; however, the mean difference in scores ("impact") for various pairs of groups have been examined (e.g., Morgan, 1992; Morgan & Maneckshana, 1992).

To determine whether application of DIF techniques to these items would be useful, an analysis was conducted of data from the 1993 Physics B exam (the less advanced of two AP Physics exams). The exam consisted of 70 multiple-choice questions and six free-response items, each of which was scored on a 0 to 15 scale. This exam was chosen in part because the multiple-choice and free-response sections are regarded as alternative measures of the same construct.

The DIF results for the physics items, which are described in the following sections, were found to yield a different perspective than does examination of the impact values. However, examination of the results by AP Physics test developers and statisticians and by another physicist did not yield a content-based explanation of the DIF findings.

5.1.1. Matching procedures

The reported score for this exam is computed by multiplying the formula score on the multiple-choice section by a factor of 1.286, and then adding the total score on the free-response items. (The goal of the AP program in choosing this weighting scheme was to assign equal weight to the multiple-choice and free-

response sections.) The SIBTEST program, however required that matching be achieved using a reliability-based correction of a number-right score; it was not designed to accommodate weights or formula scores (see Section 3.1.2). Therefore, in order to conduct a fair and straightforward comparison of SIBTEST to the other DIF methods, we did not use the reported score for matching. In the case of SIBTEST, we used the appropriately adjusted number-right score on the multiple-choice section as the matching variable. For the Mantel and SMD approaches, examinees were matched on the sum of the number-right score on the multiple-choice section and the score on the studied free-response item. That is, for each DIF procedure, matching was conducted as in the simulation study.[6] (See Section 7 for further discussion of this issue.) In our sample of about 13,000 examinees, there was a correlation of .99 between the formula score and the number-right score on the multiple-choice section.

5.1.2. Gender analysis

One comparison of the various DIF procedures was based on a DIF analysis of men and women. Analyses included only those students who stated that English was one of their best languages. The following table shows the sample sizes, along with some descriptive statistics for the multiple-choice section.

		Physics B Multiple-Choice Score		
	Sample size	Reliability	Mean	S.D.
Men	9,104	.91	37.4	12.6
Women	4,118	.89	31.1	11.4

The multiple-choice mean for men was about one-half of a standard deviation higher than the mean for women; therefore, we would expect that errors in matching of the kind described in Section 2 might occur. On the other hand, the test reliability is quite high, which is likely to be a mitigating factor.

Table 7 shows the value of the three DIF effect size measures, along with the impact values, for the six free-response items. Although the mean scores on

[6] In a typical ETS DIF analysis, a procedure called criterion refinement would have been used to screen DIF items out of the matching variable. Because the application of this procedure would not have been straightforward for SIBTEST, we did not use refinement in our comparison.

five of the six free-response items were higher for males, the DIF analyses revealed that Items 2-6 had DIF favoring females. Item 1, which had a large impact of -2.47, had DIF in the same direction--favoring males. The DIF results for all six items were statistically significant, as noted below, but only Items 5 and 6 had DIF exceeding one score point in magnitude. (Section 6.3 provides another possible way to evaluate the importance of the DIF results.)

Because the free-response scores range from 0 to 15, the DIF (and impact) measures can range from -15 to +15. It is somewhat surprising, then, that the DIF results are so similar across methods, particularly considering that in the case of the Mantel and SMD approaches, adding the studied item score to the multiple-choice score could have increased an examinee's score on the matching variable by as much as 15 points. This approach to matching might have been expected to produce results that differed substantially from those of SIBTEST, in which examinees are matched on an adjusted score on the multiple-choice items only. For Items 2, 3, 5, and 6, the DIF measures differed by no more than .02 across methods. The largest difference between any two methods occurred for Item 1, where SMD and SIBTEST differed by .44, an amount that is still only a fraction of the 30-point range of the DIF measures.

Table 8 shows the Z-statistics for the five inferential procedures included in the simulation, as well as one additional procedure. The column labeled "CMR Standardization" corresponds to a standardization DIF procedure studied by Chang, Mazzeo, and Roussos (in press). The statistic is obtained by dividing the SMD effect size by the SIBTEST standard error--it is, in fact, SIBTEST without the adjustment for measurement error. (The adjustment does not affect the SIBTEST standard error.) Comparing the CMR Standardization statistic to Standard SIBTEST provides a direct assessment of the effect of SIBTEST's measurement error correction.

At conventional significance levels, there would be no differences in conclusions across methods; all results would be significant. Mantel, SMD-H, and SMD-M, were always very similar to each other and the two SIBTEST methods always yielded similar Z-statistics. However, the Mantel and SMD methods tended to yield somewhat different results from the SIBTEST methods: On Item 1, which was the only item with negative DIF, the SIBTEST procedures yielded more extreme Z-values than the observed-score methods, while on the remaining items, the opposite was true. The CMR Standardization statistic produced results very similar to those of the SIBTEST methods except on Items 1 and 4, where its Z-values were more similar to those of the Mantel and SMD methods.

5.1.3. Analysis of Asian American and White examinees

A second comparison of the various DIF procedures was based on a DIF analysis of Asian American and White test-takers. Again, analyses included only those students who stated that English was one of their best languages. The following table shows the sample sizes, along with descriptive statistics for the multiple-choice section.

		Physics B Multiple-Choice Score		
	Sample size	Reliability	Mean	S.D.
White examinees	8,685	.91	35.4	12.4
Asian American examinees	2,364	.91	36.5	12.7

Table 9 shows the values of the three DIF effect size measures, along with the impact values, for the six free-response items; Table 10 shows the Z-statistics corresponding to the inferential procedures. The mean scores on Items 2-6 were higher for Asian American examinees. DIF analyses revealed that Item 4 had slight DIF favoring White examinees, and Items 2, 3, 6, and possibly 5 had DIF favoring Asian American test-takers. None of the items had DIF exceeding one score point in magnitude.

Again, the values of the three effect size measures were quite similar, differing here by no more than .2. SMD tended to be smaller than the SIBTEST measures. As in the gender analyses, the SMD-H, SMD-M, and Mantel Z-statistics were always very similar and the two SIBTEST methods were similar, but these two sets of procedures usually differed from each other. On Items 2, 3, 5, and 6, the SIBTEST procedures yielded more extreme Z-values than the SMD and Mantel methods, while on Item 4, the opposite was true. (Item 1 results were nearly the same for the five procedures.) The CMR Standardization statistic tended to produced smaller Z-values than the remaining methods. At conventional significance levels, only Item 5 would be judged differently across methods, with the SIBTEST statistics showing significant results while the other methods did not.

5.2. GRE writing measure

The GRE program is conducting pilot tests of a new writing measure that will become a "module" of the Graduate Record Examinations. DIF analyses were conducted on pilot test data that were collected during the fall of 1994. The examinees completed the CAT versions of the verbal, quantitative, and analytical

sections of the GRE and were then asked to respond to an essay prompt (either on the computer or by hand). There were ten prompts in all, but only one prompt was administered to each examinee. Each prompt asked the examinee to explore the implications of a statement.

According to current GRE plans, responses to all prompts are to be considered interchangeable; that is, no equating is thought to be necessary to place the prompts on a common scale. The essays were scored by two readers on a 1-6 scale. The final essay score was, in most cases, the average of the two ratings. Essays that produced large discrepancies between the two readers were subjected to a score resolution process to produce a final score.

The pilot study did not yield enough data to allow separate DIF analyses of each of the writing prompts. Because of the interchangeability assumption, it seemed reasonable to aggregate responses across prompts for DIF analysis. Responses to eight of the prompts were therefore combined; the two other prompts were deemed unsatisfactory by test development staff, based on the pilot data. Records with a "no response" or "off topic" code were deleted.

Two comparisons, described below, were conducted: one based on gender and one on language background. The Mantel and SMD methods were applied, with the matching variable defined as the sum of the Verbal CAT score (which ranged from 200 to 800) and the studied item score. The resulting variable was divided into 10-point intervals for matching. Table 11 summarizes the results of the analyses. In general, results were less similar across methods than in the AP data. The Z-statistics based on the Mantel procedure, SMD-H, and SMD-M tend to be most similar when samples are large; in the GRE application, samples were considerably smaller than in the AP and Praxis data.

5.2.1. Gender analysis

There were 262 men and 323 women (the focal group) available for this analysis after applying the usual DIF screening procedure used by the GRE program, which excludes those who do not meet certain criteria involving language, citizenship, and reason for taking the GRE. The impact, or difference in mean essay scores (female minus male) was -.05. The SMD value was .14. Although there was a reversal in sign, the DIF was not statistically significant at $\alpha = .05$ based on the SMD-H or Mantel statistics. (The SMD-M statistic slightly exceeded the critical value of 1.96.)

5.2.2. Language proficiency analysis

In this analysis, we compared those who stated that English was their best language and that they were U.S. citizens to those who failed at least one of these two criteria--the focal group. (No screening was applied prior to this analysis.) The sizes of the groups were 596 and 96, respectively. The impact was -.44; the SMD value was -.22. Although the SMD was smaller in magnitude than the impact, it was statistically significant at $\alpha = .05$ based on the Mantel and SMD-M statistics. (The Z-value of -1.93 for SMD-H was slightly smaller in magnitude than the critical value of -1.96.) This indicates that, for examinees matched on the Verbal CAT score, the average essay score for those who failed at least one of the language and citizenship criteria was lower than the average score for those who met both criteria.

5.3 Praxis I: CBT writing data

DIF analyses were conducted on essay items from Praxis I: Computer-Based Academic Skills Assessments, which is part of The Praxis Series: Professional Assessments for Beginning Teachers.™ The data were collected between the fall of 1993 and the fall of 1994. Examinees responded to a CAT writing assessment consisting of 35 multiple-choice items, followed by a single essay. There were 72 possible essay prompts; each examinee was given a choice between two of these--one in the area of "analysis of education" and one in the area of "analysis of society." Examinees had the option of writing the essay on the computer or by hand.

According to current plans, responses to all 72 prompts are to be considered interchangeable; that is, no equating is thought to be necessary to place the prompts on a common scale. The essays were scored by two readers on a 1-6 scale. Essays that produced large discrepancies between the two readers were subjected to a score resolution process to produce a final score. Records with a "no response" or "off topic" code were deleted from the DIF analyses.

Results of all Praxis analyses appear in Table 12. Three group comparisons, described below, were conducted: one comparing women and men, one comparing African American and White examinees, and one comparing those who hand-wrote their essays to those who produced their essays on the computer. Only those who stated that English was their best language were included in the analyses.[7]

---

[7] The Praxis data may have included a small number of examinees who tested under nonstandard conditions. Such examinees are not ordinarily included in operational DIF analyses.

In conducting DIF analyses, we used the mean of the two raters' scores as the Praxis essay score; therefore, our results are reported in the 1-6 metric used by the raters. This approach is consistent with our analyses of GRE data, but departs from the Praxis operational procedure of reporting the *sum* of the raters' scores. Also, when we applied the Mantel and SMD methods, our matching variable was defined as the sum of the Writing CAT score (which ranged from approximately 9 to 40) and the studied item score (in the 1-6 metric). The total scores, which are not necessarily integers, were grouped into one-unit intervals. This matching variable is not the same as the writing total score reported by Praxis, which is a combination of the CAT and essay scores which is intended to give equal weight to each of these two components.

Because of the relatively small numbers of examinees responding to each prompt, it was not possible to conduct separate analyses by prompt. Instead, responses were grouped according to the topic of the prompt, as detailed below.

5.3.1. Gender analysis of Praxis essays

For the gender comparison, the responses to the 27 education prompts were grouped together, as were responses to the 45 society prompts. As shown in Table 12, there was no evidence of DIF on these meta-items.

5.3.2. Analysis of African American and White examinees on Praxis essays

The analysis of African American and White examinees paralleled the gender comparison. In this case, Table 12 shows that there was evidence of DIF against African American test-takers on the education meta-item, with an SMD of -.13 and statistically significant Z-statistics. No DIF was evident on the society meta-item, which had an SMD value close to zero.

5.3.3. Analysis of handwritten and computer-produced Praxis essays

Somewhat over half of the examinees chose to produce their essays on the computer. These test-takers were compared to those who hand-wrote their essays--the focal group. For this comparison, sample sizes allowed a more detailed analysis than was possible for the gender and ethnic comparisons. Within the education area, prompts were coded as belonging to one of five topic areas: teacher-related (2 prompts), curriculum and evaluation (10), technology in education (2), student conduct (9), and community-related (4). Although the education and society meta-items did not show DIF, the technology in education meta-item showed DIF against those who hand-wrote their essays, a finding that

seems intuitively reasonable. The SMD was -.21, the largest of those reported in Table 12, and results were statistically significant.[8]

5.4. Interpretation of DIF results for ETS data

Several limitations should be kept in mind when interpreting the DIF results described in Sections 5.1 through 5.3. First, the statistical methods that were applied are being explored for research purposes and have not been officially adopted by ETS. Second, in the case of Praxis and GRE, the responses to several prompts were combined for analysis. This aggregation was done to obtain sufficiently large samples for illustrating the DIF methods, but is not a standard approach in DIF assessment. Third, the determination of an appropriate matching variable for these analyses was not straightforward (e.g., see Sections 5.1.1 and 7); other choices could have led to different results. The polytomous items and their corresponding matching tests may have measured somewhat different constructs, particularly in the case of the GRE and Praxis analyses. This in itself can produce DIF. Fourth, as in any DIF analysis, potential confounding variables exist. For example, the Praxis program points out that there are two groups of Praxis test-takers: those who are entering teacher education programs and those who are being licensed at the end of their training. Membership in these groups may be confounded with the grouping variables used in the DIF analyses (e.g., ethnicity and gender), which could complicate the interpretation of DIF findings. Another difficulty in interpreting the Praxis results is that examinees were allowed to select either an education or a society essay prompt.

The determination of which results should be judged as large is also complex. As is always the case, statistical significance does not imply practical importance. All the DIF measures reported here are in the score point metric; DIF that is a small fraction of a score point in magnitude may not be worrisome even if it is statistically significant. Section 6 below offers possible criteria for judging importance based on both statistical significance and the size of the DIF index; application of the criteria to the ETS data sets is discussed in Section 6.3.

6. DIF categorization procedures

The goal of this portion of the study was to take some initial steps toward developing a system for classifying polytomous items which was analogous to the

---

[8] Interpretation of the differences in results for handwritten versus computer-produced essays is complicated by the fact that handwritten essays were not key-entered before scoring. This may in itself have affected raters. However, because there is no reason to expect the effect on raters to differ across the essay topics, the finding on the technology meta-item seems unlikely to be the result of a rater bias of this kind.

ETS system for characterizing dichotomous items as having no DIF ("A"), moderate DIF ("B"), or large DIF ("C"). A system of this kind would classify items on the basis of both a hypothesis test and a measure of DIF effect size. In developing such a system for polytomous items, it is necessary to take into account the fact that measures of DIF in polytomous items like those in Section 3.1 are expressed in the item score metric. That is, the size of the DIF measure is meaningful only with reference to the scale used for scoring the item (e.g., 1-6). Although this property is useful for some purposes, it is an obstacle to the development of a classification rule that can be applied to all polytomous items, regardless of score scale. One possible way to eliminate this scale-dependence property is to begin with a DIF measure in the item score metric and then divide by a measure of item score variability, such as the standard deviation or range. This is intended to yield a DIF effect size index that is *scale-invariant* (i.e., independent of the score metric of the item) and can therefore be compared across items with different score scales. In the polytomous case, then, the proposed rules would classify an item's DIF severity using (a) the results of one of the hypothesis testing procedures in Section 3.2 and (b) the magnitude of a ratio statistic obtained by dividing a measure of DIF effect size from Section 3.1 by a measure of item score variability. Our work in this area, which should be regarded as preliminary, is described in the following sections.

First, we attempted to generalize the rules used in evaluating the size of STD P-DIF (Dorans & Kulick, 1986) to items with more than two response categories, using a ratio measure of the kind described above. For the case of dichotomous items, we then compared the results of applying the newly developed general rules to the results of applying the current STD P-DIF rules.

Second, we evaluated the degree to which the proposed ratio measures were invariant to the item score scale by comparing results obtained for simulated dichotomous and polytomous items with the same value of d (the difference between reference and focal difficulties). In conducting this work, we considered the three DIF indexes of Section 3.1 as possible numerators for the ratio statistics and also considered several measures of item variability as candidates for the denominator. Unless otherwise noted, the results reported here are for the ratio of SMD to a pooled item standard deviation, which was obtained as follows for item i:

$$\hat{\sigma}_i = \sqrt{\frac{(n_R - 1)\hat{\sigma}_{iR}^2 + (n_F - 1)\hat{\sigma}_{iF}^2}{(n_R + n_F - 2)}}, \tag{5}$$

where $\hat{\sigma}_{iR}^2$ and $\hat{\sigma}_{iF}^2$ are the item i variances in the reference and focal groups, respectively. The results of Section 4.4 suggest that, in future applications, it

may be preferable to use the Modified SIBTEST measure, rather than SMD, as a numerator.

Third, we applied the proposed rules to the ETS data sets analyzed in Section 5. We also report the results of an application to data from the National Assessment of Educational Progress (NAEP).

6.1. Generalization of STD P-DIF rules

The goal of the analyses described in this section was to produce rules for classifying polytomous item into A, B, and C categories that were, in some sense, equal in stringency to the rules used in the dichotomous case. For STD P-DIF (which is identical to SMD in dichotomous items), magnitudes over .05 are considered worthy of examination and values over .10 are considered substantial (Dorans & Kulick, 1986). If most item proportions correct ($\hat{p}$) range from .1 to .9, then most item standard deviations ($\sqrt{\hat{p}(1-\hat{p})}$) range from .3 to .5. Taking .4 as a typical item standard deviation value suggests that cut-offs of .05 and .10 for STD P-DIF correspond to cut–offs of .125 and .25 for SMD/item SD. We tested out the agreement between these rules for 18 simulated items. The items were created to correspond to the studied items in the simulation of Section 4, except that they had only two possible response categories. That is, the a, b_R, and d values for the items were the same as those listed in Table 1, and the model used for data generation was a dichotomous version of the model in equation 3. (It was assumed that there was no guessing.) The matching test consisted of fifty dichotomous items. For each item and each of the two focal group conditions, 250 replications were conducted. The two rules for classifying DIF were as follows:

Current STD P-DIF Rule

If STD P-DIF$_i$ < -.10, I = 1
If -.10 < STD P-DIF$_i$ < -.05, I = 2
If -.05 < STD P-DIF$_i$ < .05, I = 3
If .05 < STD P-DIF$_i$ < .10, I = 4
If STD P-DIF$_i$ > .10, I = 5

Generalized Rule

If STD P-DIF$_i$ / $\hat{\sigma}_i$ < -.25, J = 1
If -.25 < STD P-DIF$_i$ / $\hat{\sigma}_i$ < -.125, J = 2
If -.125 < STD P-DIF$_i$ / $\hat{\sigma}_i$ < .125, J = 3
If .125 < STD P-DIF$_i$ / $\hat{\sigma}_i$ < .25, J = 4
If STD P-DIF$_i$ / $\hat{\sigma}_i$ > .25, J = 5

The agreement between classifications I and J was assessed using the proportions of exact agreement (over the 250 replications), as well as kappa statistics, which are "chance-corrected" measures of agreement (Cohen, 1960). The ranges and medians of these summary statistics over the 18 items were as follows:

	Proportion Agreement		Kappa	
Focal condition	Range	Median	Range	Median
N(0, 1)	.69 to .99	.83	.29 to .80	.63
N(-1, 1)	.72 to 1.00	.88	.50 to 1.00	.73

Agreement was moderately good, but far from perfect. Obviously, the degree of agreement depends in part on the assumption that was made about the size of a typical item standard deviation in the dichotomous case. It is not clear why performance was better for the N(-1, 1) focal group condition than for the N(0, 1) condition.

6.2. Comparison of ratio statistics for dichotomous and polytomous items

A second analysis was conducted to determine whether SMD/ $\hat{\sigma}_i$ would behave similarly, as desired, for four-category items as for analogous dichotomous items. Results for the 18 polytomous items of Table 1 were compared to those for their dichotomous counterparts, described above. Figures 1 and 2 show, for the N(0, 1) and N(-1, 1) focal group conditions, respectively, the means (over 250 replications) of these ratio statistics for the polytomous items plotted against the corresponding values for the dichotomous items. Tables 13 and 14 show, for the N(0, 1) and N(-1, 1) focal group conditions, respectively, the ratio statistics for the polytomous (column 2) and dichotomous (column 3) items, along with the deviations (polytomous minus dichotomous, column 4) and absolute deviations (column 5). Each table is ordered by the size of the absolute deviation. For the N(0, 1) focal group condition, the absolute deviations ranged from 0 to .054, with a median of .026. Absolute deviations were smallest for items with no DIF. For the N(-1, 1) condition, the absolute deviations ranged from 0 to .059, with a median of .025. Absolute deviations were largest for items with negative DIF. Again, performance of these ratios was acceptable, but not ideal.

6.3. Application of proposed rules to actual data

In attempting to parallel the current system for classifying items into A, B, and C categories, we could consider a rule in which absolute values of SMD/$\hat{\sigma}_i$ that exceed .125 are indicative of "B" status, provided that results are statistically significant at α = .05, and absolute values that exceed .25 indicate "C" status, again provided that results are statistically significant. Application of this rule leads to the following results for the ETS analyses reported in Section 5: In the gender analysis of AP data, Items 2 and 3 are B items, with SMD/$\hat{\sigma}_i$ values of .22 and .18, and Items 5 and 6 are C items, with SMD/$\hat{\sigma}_i$ values of .36 and .28. In the Asian American-White comparison, Item 6 is a B, with an SMD/$\hat{\sigma}_i$ value of .14. All these B and C items show DIF in favor of the focal groups. The GRE language comparison yields a B result (SMD/$\hat{\sigma}_i$ = -.21) and the comparison of handwritten and computer-produced Praxis essays leads to a C result (SMD/$\hat{\sigma}_i$ = -.27); in these cases, the DIF is in favor of the reference groups. All other results fail to meet these (very tentative) B and C criteria.

Recently, NAEP applied a rule based on a statistic similar to SMD/$\hat{\sigma}_i$, along with the Mantel Z test, for purposes of screening polytomous items for DIF. (The NAEP standard deviation was based on the reference and focal groups combined, whereas in equation 5 above, $\hat{\sigma}_i$ is defined as the square root of the pooled *within-group* variance. Also, NAEP did not use cut-offs of .125 and .25 for determining which items had DIF. These criteria are applied here for illustrative purposes.) The 1994 NAEP reading assessment (Williams, Reese, Campbell, Mazzeo, & Phillips, 1995) included 17 polytomous items at grade 4, 32 such items at grade 8, and 35 such items at grade 12. At each grade level, the comparisons that were conducted were Female-Male, African American-White, and Hispanic-White, for a total of 252 item-level DIF analyses. There were no items in any comparison that had both a significant Mantel statistic (α = .05) and a ratio value that exceeded .25 in magnitude. Four items had both significant Mantel statistics and ratios exceeding .125--one at grade 4 (favoring African American over White examinees), one at grade 8 (favoring White over African American examinees), and two at grade 12 (one favoring Hispanic over White examinees and one in the opposite direction).

7. Summary and discussion

Our study of DIF methods for polytomous items included several components: a large-scale simulation that evaluated three descriptive statistics and five inferential procedures, applications of the various statistical methods to three ETS data sets, and initial work on the development of a system for categorizing DIF severity in polytomous items which is analogous to the ETS

system now in place for dichotomous items. In addition, two alternative standard error formulas were developed for SMD, a descriptive DIF statistic that is a generalization of the STD P-DIF index now used at ETS for dichotomous items; this work is documented in Zwick and Thayer (1994, in press).

The simulation study showed that, when the reference and focal groups had identical distributions (i.e., the N(0, 1) focal group condition), SMD performed best as a descriptive statistic, but when the focal group mean was one standard deviation lower than the reference group mean (i.e., the N(-1, 1) focal group condition), Modified SIBTEST tended to produce the best results.

In the N(0, 1) focal group condition, the five inferential procedures performed almost indistinguishably. However, when the focal group had a N(-1, 1) distribution and the discrimination of the studied item was higher than the average discrimination of the matching items, the two SIBTEST procedures showed much better Type I error control than the SMD and Mantel methods, which showed very high error rates. Test length had a substantial modifying influence on this Type I error inflation. As previously found, SMD-H provided somewhat better Type I error control than SMD-M.

The power ranking of the five procedures was inconsistent; it depended on the direction of DIF and other factors. Modified SIBTEST behaved in a more symmetric way than its competitors in that its rejection rates for $d = -.25$ were similar to its rejection rates for $d = .25$. Overall, the descriptive and inferential performance of Modified SIBTEST appeared to be superior to that of SIBTEST.

Application of all the DIF methods to data from the AP Physics exam showed remarkably similar results across methods. In general, the same conclusions would have been drawn from any of the possible approaches. A practical drawback of SIBTEST in its current version is that it cannot accommodate matching variables that are obtained through scoring algorithms other than number-right scoring. Therefore, SIBTEST was not included in the applications to GRE and Praxis data. For the Praxis data, the Mantel method and the two SMD procedures tended to show similar results; this was less true for the GRE data, perhaps because of smaller samples.

The work completed for this project suggested a number of areas in which further research is likely to be valuable. First, more work is needed to understand exactly what conditions tend to produce the best and worst performance for DIF methods that match on observed score, such as the Mantel and SMD procedures. Theory tells us that these procedures perform best in conditions that resemble the Rasch case, but even in a simulation, it is not straightforward to determine which conditions most closely resemble the Rasch

case. In fact, our conjecture about the condition that would optimize the performance of the observed-score methods proved to be wrong. Developing an index of departure from "Raschness," as proposed by Holland, may be useful.

A related issue concerns the inclusion of the studied item in the matching variable in the observed-score methods. If all items in the test follow the Rasch or partial-credit models, there is a theoretical justification for adding in the studied item with no rescaling (i.e., if the item is scored 1-15, the score should not be, say, divided by 15 before adding it in). Under these special models, the rationale for the no-rescaling policy is that the total test score (with no rescaling) is a sufficient statistic for ability (see Section 2 and Zwick, Donoghue, & Grima, 1993a, 1993b). However, under other models, it is not clear that it is best to avoid rescaling. Perhaps a rescaled item score would more closely follow the partial credit model than the unrescaled score in some cases. In any event, the results obtained using the observed-score methods do depend on whether rescaling is used. Similarly, the observed-score DIF methods would not necessarily lead to the same conclusions if the sum of two raters' scores were analyzed in place of the mean of the two raters' scores. By contrast, SIBTEST does lead to the same conclusion in these two cases.

Initial work on the development of a system for categorizing an item's DIF severity as A, B, or C, as is now done in the dichotomous case, led to useful, but incomplete results. A ratio statistic obtained by dividing a DIF effect size measure by the item standard deviation was evaluated as a possible DIF measure that would be independent of the score scale of the item. Results were somewhat promising, but more work is needed. Part of the difficulty in developing a measure that is invariant to the score scale lies in the intractability of the item score metric. The difference between groups in expected item score is not constant over ability, even under simple item response models.

Finally, a remaining unresolved issue is the development of DIF methods or equivalent procedures for performance tasks or other polytomous items for which no appropriate multiple-choice matching variable is available. One possibility is to use a multivariate matching procedure based on all available cognitive and background variables. A more promising approach may be to focus on differential validity rather than DIF. The goal would then be to compare the relationship between the item score and a criterion measure for the demographic groups of interest. If performance assessment continues to grow at the present rate, the challenge of conceptualizing and assessing the fairness of performance tasks will become increasingly important.

References

Allen, N. L., & Donoghue, J. R. (in press). Applying the Mantel-Haenszel procedure to complex samples of items. *Journal of Educational Measurement..*

Chang, H., Mazzeo, J., & Roussos, L. (in press). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *Journal of Educational Measurement.*

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37-46.

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 and H. Wainer (Eds.), *Differential Item Functioning.* Hillsdale, NJ: Erlbaum.

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. (1991). *Constructed response and differential item functioning: A pragmatic approach.* (ETS Research Report 91-47.) Princeton, NJ: Educational Testing Service.

Douglas, J., Stout, W., & DiBello, L. (1995). *A kernel smoothed version of SIBTEST with applications to local DIF inference and function estimation.* Unpublished paper. Department of Statistics, University of Illinois, Champaign.

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

Holland, P. W. (January 14, 1991). *Item and DIF analyses for items with ordered responses.* Internal ETS memorandum.

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

Kelley, T. L. (1923). *Statistical methods.* New York: Macmillan.

Lewis, C. (1993). A note on the value of including the studied item in the test score when analyzing test items for DIF. In Holland, P. W., & Wainer, H. (Eds.), *Differential Item Functioning*, pp. 317-319. 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. (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.

Mazzeo, J., & Chang, H.-H. (1994, April). *Detecting DIF for polytomously scored items: Progress in adaptation of Shealy-Stout's SIBTEST procedure.* Presented at the annual meeting of the American Educational Research Association, New Orleans.

Meredith, W., & Millsap, R. E. (1992). On the misuse of manifest variables in the detection of measurement bias. *Psychometrika, 57,* 289-311.

Millsap, R. E., & Meredith, W. (1992). Inferential conditions in the statistical detection of measurement bias. *Applied Psychological Measurement, 16,* 389-402.

Morgan, R. (May 20, 1992). *Subgroup performance of AP free response items.* Internal memorandum, Educational Testing Service.

Morgan, R., & Maneckshana, B. (March 5, 1992). *Subgroup reliability for the Advanced Placement English Language and Composition, European History, and United States Government and Politics Examinations (Form 3NBP).* ETS memorandum.

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

Potenza, M. T, & Dorans, N. J. (1995). DIF assessment for polytomously scored items: A framework for classification and evaluation. *Applied Psychological Measurement, 19,* 23-37.

Roussos, L. A., & Stout, W. F. (1993, April). *Simulation studies of effects of small sample sized and studied item parameters on SIBTEST and Mantel-Haenszel Type I error performance.* Presented at the annual meeting of the American Educational Research Association, Atlanta.

Shealy, R. (1989). *An item response theory-based statistical procedure for detecting concurrent internal bias in ability tests.* Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.

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.

Uttaro, T., & Millsap, R. E. (1994). Factors influencing the Mantel-Haenszel procedure in the detection of differential item functioning. *Applied Psychological Measurement, 18,* 15-25.

Welch, C. J., & Miller. T. (1995). Assessing differential item functioning in direct writing assessments: Problems and an example. *Journal of Educational Measurement, 32* 163-178.

Williams, P. L., Reese, C. M., Campbell, J., Mazzeo, J., & Phillips, G. W. (1995). *1994 NAEP Reading: A First Look.* NCES Report No. 95-748. Washington, DC: National Center for Education Statistics.

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. (1992). *Application of Mantel's chi-square test to the analysis of differential item functioning for ordinal items.* Technical memorandum.

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

Zwick, R., Donoghue, J. R., & Grima, A. (1993b). *Assessing differential item functioning in performance tests.* (ETS Research Report 93-14). Princeton, NJ: Educational Testing Service.

Zwick, R., & Thayer, D. T. (1994). *Evaluation of the magnitude of differential item functioning in polytomous items* (ETS Research Report 94-13.) Princeton, NJ: Educational Testing Service.

Zwick, R., & Thayer, D. T. (July 12, 1995). *Standard errors for the STD P-DIF statistic.* ETS memorandum.

Zwick, R., & Thayer, D. T. (in press). Evaluating the magnitude of differential item functioning in polytomous items. *Journal of Educational and Behavioral Statistics.*

Zwick, R., Thayer, D. T., and Wingersky, M. (1994). A simulation study of methods for assessing differential item functioning in computerized adaptive tests. *Applied Psychological Measurement, 18,* 121-140.

Table 1

Item Parameters and True DIF Values

				True DIF Values	
Item	a	b_R	d	Focal N(0, 1)	Focal N(-1, 1)
1	.47	-.5	.25	.145	.154
2	.47	-.5	0	0	0
3	.47	-.5	-.25	-.151	-.146
4	.47	.5	.25	.151	.115
5	.47	.5	0	0	0
6	.47	.5	-.25	-.145	-.103
7	.86	-.5	.25	.179	.190
8	.86	-.5	0	0	0
9	.86	-.5	-.25	-.190	-.179
10	.86	.5	.25	.190	.128
11	.86	.5	0	0	0
12	.86	.5	-.25	-.179	-.108
13	1.57	-.5	.25	.201	.216
14	1.57	-.5	0	0	0
15	1.57	-.5	-.25	-.216	-.201
16	1.57	.5	.25	.216	.132
17	1.57	.5	0	0	0
18	1.57	.5	-.25	-.201	-.107

Note. a is the studied item discrimination, b_R is the reference group difficulty, and d is the difference between the reference and focal group difficulties.

Table 2

Average Residuals (DIF Measure - True DIF) for Each Level of Discrimination (a), Difference in Reference and Focal Difficulty (d), and Focal Group Distribution

Focal Group	N(0, 1)			N(-1, 1)		
a	.47	.86	1.57	.47	.86	1.57
$d = -.25$						
SMD	-.007	-.023	-.026	.040	.017	.009
Std. SIBTEST	-.024	-.049	-.058	-.001	.011	.023
Mod. SIBTEST	-.024	-.049	-.058	.014	.031	.047
$d = 0$						
SMD	-.002	-.001	.000	.007	-.031	-.059
Std. SIBTEST	-.002	-.001	-.000	-.020	-.016	-.018
Mod. SIBTEST	-.002	-.001	-.000	-.005	.004	.005
$d = .25$						
SMD	-.003	.001	.007	-.015	-.069	-.107
Std. SIBTEST	.013	.025	.039	-.027	-.032	-.037
Mod. SIBTEST	.013	.025	.039	-.013	-.012	-.014

Note. Each entry is averaged over the values of b_R and matching test length and is based on 2,000 observations. The standard errors are approximately .001 for most entries. For the N(-1, 1) condition, $d = \pm .25$, some standard errors are slightly larger, but none exceed .002.

Table 3

Average Residuals (DIF Measure - True DIF) for Each Level of Reference Group Difficulty (b_R), Difference in Reference and Focal Difficulty (d), and Focal Group Distribution

Focal Group	N (0, 1)		N (-1, 1)	
b_R	-.5	.5	-.5	.5
d = -.25				
SMD	.010	-.048	-.013	.057
Std. SIBTEST	-.015	-.072	-.026	.048
Mod. SIBTEST	-.015	-.072	-.008	.068
d = 0				
SMD	.001	-.002	-.037	-.018
Std. SIBTEST	.001	-.003	-.026	-.011
Mod. SIBTEST	.001	-.003	-.008	.011
d =.25				
SMD	-.035	.038	-.042	-.085
Std. SIBTEST	-.011	.063	-.004	-.060
Mod. SIBTEST	-.011	.063	.013	-.038

Note. Each entry is averaged over the values of a and matching test length and is based on 3,000 observations. The standard errors are less than .001.

Table 4

Statistically Significant Predictors of Rejection Rates

Effect	SMD-H	SMD-M	Mantel	Std. SIBTEST	Mod. SIBTEST
		No-DIF Items Only			
A x F	•	•	•		
A x B x F	•	•	•		
A x F x L	•	•	•		
		All Items Simultaneously			
A	•	•	•	•	•
D	•	•	•	•	•
F				•	•
A x B	•	•	•	•	•
A x D	•	•	•	•	•
A x F	•	•	•	•	
A x L	•	•	•	•	•
D x F			•	•	
A x B x D	•	•		•	•
A x D x F	•	•	•	•	•
B x D x L	•	•	•	•	•
B x D x F					•
A x F x L		•			

Note. A dot indicates that the corresponding effect was statistically significant ($p < .005$) for the indicated procedure. A, B, and D are the discrimination, reference group difficulty, and difference between reference and focal difficulty, respectively, of the studied item. F is the focal group distribution and L is test length.

Table 5
Rejection Rates for Each Level of Discrimination (a), Difference in Reference and Focal Difficulty (d), and Focal Group Distribution

	Studied Item Discrimination (a)		
	0.47	0.86	1.57
d = -0.25			
N(0,1) Focal Group			
SMD-H	0.71	0.96	0.99
SMD-M	0.72	0.96	0.99
Mantel	0.72	0.96	0.99
Standard SIBTEST	0.73	0.96	1.00
Modified SIBTEST	0.73	0.96	0.99
N(-1,1) Focal Group			
SMD-H	0.48	0.95	1.00
SMD-M	0.50	0.96	1.00
Mantel	0.53	0.97	1.00
Standard SIBTEST	0.68	0.91	0.98
Modified SIBTEST	0.58	0.81	0.91
d = 0.0			
N(0,1) Focal Group			
SMD-H	0.05	0.05	0.05
SMD-M	0.06	0.06	0.06
Mantel	0.05	0.05	0.05
Standard SIBTEST	0.06	0.06	0.06
Modified SIBTEST	0.06	0.06	0.06
N(-1,1) Focal Group			
SMD-H	0.04	0.12	0.35
SMD-M	0.05	0.13	0.35
Mantel	0.04	0.12	0.34
Standard SIBTEST	0.07	0.09	0.11
Modified SIBTEST	0.07	0.09	0.13
d = 0.25			
N(0,1) Focal Group			
SMD-H	0.69	0.94	1.00
SMD-M	0.72	0.95	1.00
Mantel	0.70	0.95	1.00
Standard SIBTEST	0.72	0.95	1.00
Modified SIBTEST	0.72	0.95	1.00
N(-1,1) Focal Group			
SMD-H	0.62	0.65	0.70
SMD-M	0.65	0.69	0.72
Mantel	0.72	0.74	0.76
Standard SIBTEST	0.49	0.79	0.96
Modified SIBTEST	0.57	0.87	0.98

Note. Each entry is averaged over the values of b_R and matching test length and is based on 2,000 observations. Standard errors range from .001 to .011.

Table 6
Rejection Rates for Each level of Discrimination (a), Difference in Reference and Focal Difficulty (d), and Matching Test Length for the N(-1, 1) Focal Group Condition

	Studied Item Discrimination (a)		
	0.47	0.86	1.57
d = -0.25			
20 Matching Items			
SMD-H	0.46	0.97	1.00
SMD-M	0.46	0.97	1.00
Mantel	0.49	0.98	1.00
Standard SIBTEST	0.69	0.89	0.97
Modified SIBTEST	0.54	0.74	0.85
50 Matching Items			
SMD-H	0.51	0.93	1.00
SMD-M	0.54	0.94	1.00
Mantel	0.57	0.96	1.00
Standard SIBTEST	0.67	0.92	0.99
Modified SIBTEST	0.63	0.88	0.97
d = 0.0			
20 Matching Items			
SMD-H	0.05	0.15	0.50
SMD-M	0.06	0.16	0.49
Mantel	0.05	0.15	0.49
Standard SIBTEST	0.08	0.10	0.12
Modified SIBTEST	0.08	0.10	0.16
50 Matching Items			
SMD-H	0.03	0.08	0.20
SMD-M	0.05	0.09	0.22
Mantel	0.04	0.08	0.19
Standard SIBTEST	0.06	0.08	0.10
Modified SIBTEST	0.07	0.08	0.09
d = 0.25			
20 Matching Items			
SMD-H	0.64	0.56	0.50
SMD-M	0.66	0.59	0.53
Mantel	0.74	0.65	0.58
Standard SIBTEST	0.46	0.75	0.93
Modified SIBTEST	0.58	0.87	0.98
50 Matching Items			
SMD-H	0.59	0.74	0.89
SMD-M	0.65	0.79	0.90
Mantel	0.70	0.84	0.94
Standard SIBTEST	0.52	0.83	0.98
Modified SIBTEST	0.56	0.87	0.99

Note. Each entry is averaged over the values of b_R and is based on 1,000 observations. Standard errors range from .001 to .016.

Table 7

Values of Impact and DIF Effect Size for Gender Analysis of AP Physics B Free-Response Items

Item	Impact	SMD	Standard SIBTEST	Modified SIBTEST
1.	-2.47	-.51	-.95	-.93
2.	-0.58	.97	.96	.98
3.	-0.60	.68	.67	.69
4.	-1.39	.44	.26	.29
5.	0.31	1.54	1.54	1.55
6.	-0.11	1.09	1.08	1.10

Note. A negative impact statistic indicates a lower mean item score for women than for men. A negative DIF measure indicates a lower mean item score for women than for men, conditional on the matching variable.

Table 8

Z-Values for Gender DIF Analysis of
AP Physics B Free-Response Items

Item	SMD-H	SMD-M	Mantel	CMR Standardization	Standard SIBTEST	Modified SIBTEST
1.	-9.32	-9.46	-9.89	-7.73	-14.35	-14.09
2.	18.75	19.32	18.81	15.65	15.43	15.80
3.	15.47	15.54	15.56	12.83	12.54	13.02
4.	8.08	8.26	7.94	6.47	3.86	4.24
5.	27.47	28.40	27.79	23.33	23.26	23.50
6.	23.73	23.73	23.14	19.82	19.64	20.10

Note. A negative DIF statistic indicates a lower mean item score for women than for men, conditional on the matching variable. The CMR Standardization statistic is obtained by dividing SMD by the SIBTEST standard error.

Table 9

Values of Impact and DIF Effect Size for Asian American-White Analysis of AP Physics B Free-Response Items

Item	Impact	SMD	Standard SIBTEST	Modified SIBTEST
1.	-0.15	-0.51	-0.56	-0.57
2.	0.97	0.39	0.58	0.57
3.	0.80	0.34	0.47	0.46
4.	0.38	-0.10	-0.07	-0.08
5.	0.56	0.13	0.25	0.25
6.	1.03	0.54	0.75	0.74

Note. A negative impact statistic indicates a lower mean item score for Asian American examinees than for White examinees. A negative DIF measure indicates a lower mean item score for Asian American than for White examinees, conditional on the matching variable.

Table 10

Z-Values for Asian American-White DIF Analysis of AP Physics B Free-Response Items

Item	SMD-H	SMD-M	Mantel	CMR Standardization	Standard SIBTEST	Modified SIBTEST
1.	-7.69	-7.87	-7.59	-6.38	-7.04	-7.12
2.	5.92	6.04	6.08	4.88	7.32	7.23
3.	5.84	5.81	5.92	4.86	6.59	6.47
4.	-1.43	-1.44	-1.38	-1.11	-0.80	-0.89
5.	1.70	1.73	1.76	1.44	2.93	2.85
6.	8.23	8.08	8.35	6.75	9.09	8.97

Note. A negative DIF statistic indicates a lower mean item score for Asian-American examinees than for White examinees, conditional on the matching variable. The CMR Standardization statistic is obtained by dividing SMD by the SIBTEST standard error.

Table 11

GRE Writing Pilot Data

	Descriptive Results		Z - Statistics		
Comparison	Impact	SMD	SMD-H	SMD-M	Mantel
Gender[a]	-0.05	0.14	1.62	2.05	1.28
Language[b]	-0.44	-0.22	-1.93	-2.72	-2.39

[a]The analysis compared 262 men to 323 women. A negative impact statistic indicates a lower mean item score for women than for men. A negative DIF measure indicates a lower mean item score for women than for men, conditional on the matching variable.

[b]The analysis compared 596 examinees who met language/citizenship criteria to 96 who did not . A negative impact statistic indicates a lower mean item score for those who did not meet the criteria than for those who did. A negative DIF measure indicates a lower mean item score for those who did not meet the criteria than for those who did, conditional on the matching variable.

Table 12

Results of DIF Analysis of Praxis I Essays

Topic	Sample Size R	Sample Size F	Descriptive Results Impact	Descriptive Results SMD	Z-Statistics SMD-H	Z-Statistics SMD-M	Z-Statistics Mantel
Gender Comparison							
Education (combined)	809	1,621	.11	.02	.55	.57	.95
Society (combined)	491	825	.07	.05	1.09	1.15	1.11
African American-White Comparison							
Education (combined)	2,025	261	-.47	-.13	-2.76	-2.78	-2.65
Society (combined)	1,077	170	-.44	-.05	-0.73	-0.88	-0.31
Handwritten Essay - Computer Essay Comparison							
Education (combined)	1,354	1,076	-.26	-.04	-1.32	-1.36	-1.27
Teacher related	155	104	-.52	-.15	-1.52	-1.95	-1.84
Curriculum & evaluation	622	534	-.26	-.04	-1.00	-1.06	-1.27
Technology in education	130	81	-.43	-.21	-2.09	-2.69	-2.04
Student conduct	292	220	-.09	.12	1.86	2.01	1.83
Community related	185	137	-.16	.06	0.68	0.76	0.52
Society (combined)	783	563	-.28	.00	0.11	0.11	0.29

Note. A negative impact statistic indicates a lower mean item score for the focal group (women, African American examinees, or examinees who wrote essays by hand) than for the reference group (men, White examinees, or examinees who produced essays on the computer). A negative DIF measure indicates a lower mean item score for the focal group than for the reference group, conditional on the matching variable.

Table 13

Average Ratio of SMD to Item Standard Deviation for Polytomous and Dichotomous Items: N(0, 1) Focal Group Condition

Item	$SMD_i/\hat{\sigma}_i$			
	Polytomous	Dichotomous	Deviation	Absolute Deviation
8	-0.001	-0.001	0.000	0.000
14	0.001	-0.001	0.001	0.001
11	0.001	-0.002	0.003	0.003
17	0.002	-0.002	0.004	0.004
2	-0.000	0.004	-0.005	0.005
5	-0.008	0.006	-0.014	0.015
18	-0.181	-0.162	-0.019	0.019
15	-0.187	-0.166	-0.020	0.020
16	0.191	0.168	0.023	0.023
7	0.159	0.131	0.028	0.028
13	0.185	0.153	0.032	0.032
1	0.128	0.094	0.034	0.034
10	0.170	0.131	0.039	0.039
6	-0.131	-0.090	-0.040	0.040
12	-0.169	-0.124	-0.045	0.045
3	-0.131	-0.084	-0.047	0.047
4	0.135	0.087	0.049	0.049
9	-0.177	-0.123	-0.054	0.054

Note. Each ratio is averaged over 250 replications.

Table 14

Average Ratio of SMD to Item Standard Deviation for Polytomous and Dichotomous Items: N(-1, 1) Focal Group Condition

	Value of $SMD_i/\hat{\sigma}_i$			
Item	Polytomous	Dichotomous	Deviation	Absolute Deviation
16	0.113	0.112	0.000	0.000
10	0.115	0.113	0.002	0.002
7	0.145	0.149	-0.004	0.004
17	-0.029	-0.017	-0.012	0.012
11	-0.015	-0.001	-0.014	0.014
2	0.007	0.022	-0.015	0.015
5	-0.000	0.018	-0.018	0.018
4	0.121	0.102	0.018	0.019
1	0.141	0.119	0.022	0.022
8	-0.027	0.000	-0.027	0.027
15	-0.247	-0.219	-0.028	0.028
13	0.139	0.170	-0.031	0.031
18	-0.154	-0.117	-0.036	0.036
12	-0.143	-0.103	-0.041	0.041
14	-0.058	-0.016	-0.042	0.042
3	-0.129	-0.076	-0.053	0.053
6	-0.110	-0.056	-0.053	0.053
9	-0.206	-0.148	-0.059	0.059

Note. Each ratio is averaged over 250 replications.

Figure 1
Ratio of SMD to Item Standard Deviation for Polytomous and Dichotomous Items (averaged over 250 replications): N(0,1) Focal Group Condition

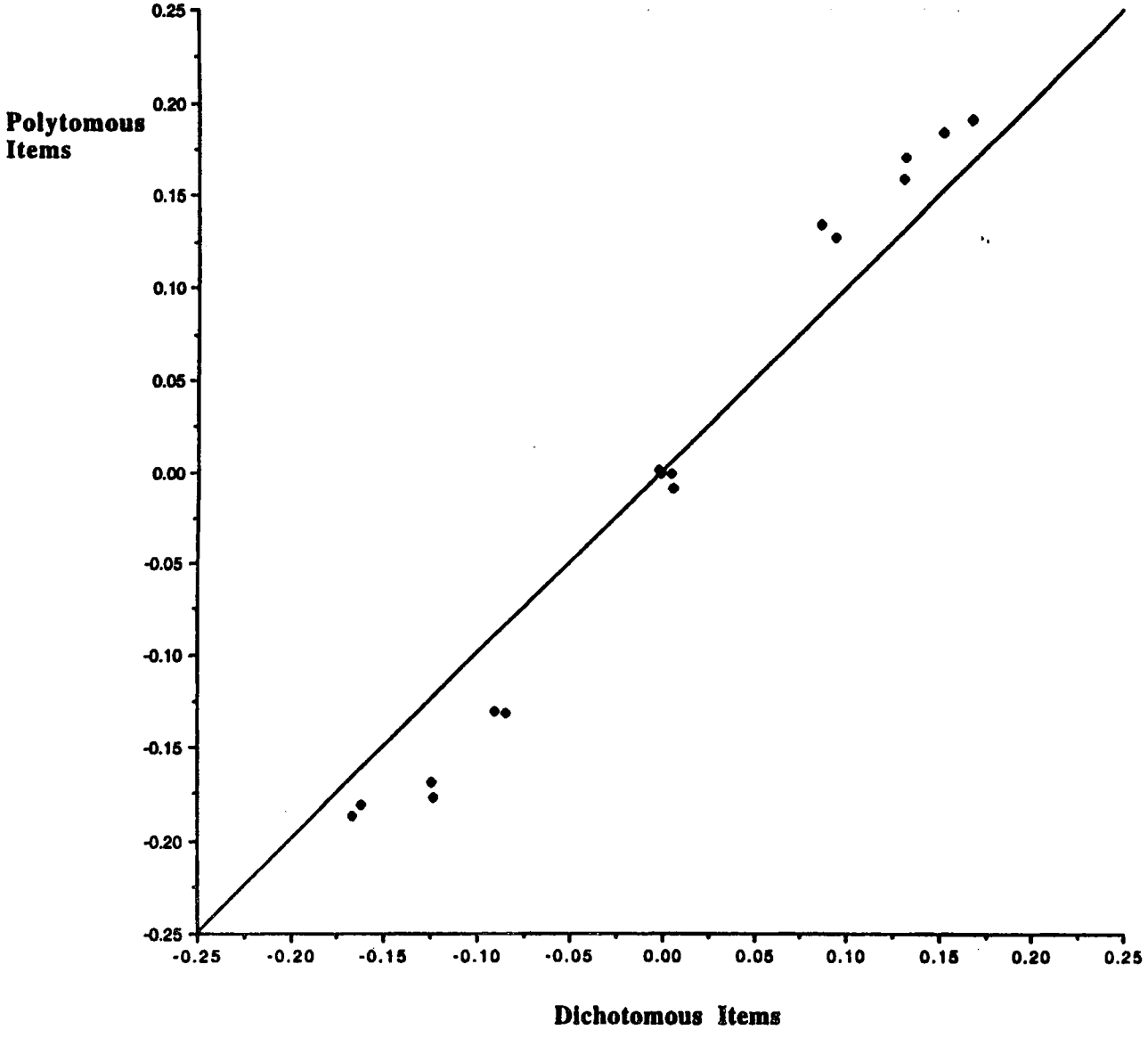

Figure 2
Ratio of SMD to Item Standard Deviation for Polytomous and Dichotomous Items (averaged over 250 replications): N(-1,1) Focal Group Condition