DOCUMENT RESUME

ED 382 666 TM 023 104

AUTHOR Zwick, Rebecca; Thayer, Dorothy T.
TITLE Evaluation of the Magnitude of Differential Item Functioning in Polytomous Items. Program Statistics Research Technical Report No. 94-2.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-94-13
PUB DATE Mar 94
NOTE 29p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 4-8, 1994).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Error of Measurement; *Evaluation Methods; *Hypothesis Testing; *Item Bias; Scaling; Scoring; Simulation; *Statistical Inference; Test Items
IDENTIFIERS Mantel Haenszel Procedure; *Polytomous Items

ABSTRACT

Several recent studies have investigated the application of statistical inference procedures to the analysis of differential item functioning (DIF) in test items that are scored on an ordinal scale. Mantel's extension of the Mantel-Haenszel test is a possible hypothesis-testing method for this purpose. The development of descriptive statistics for characterizing DIF in polytomous test items has received less attention. A statistic that appears well-suited as a summary index was proposed by Dorans and Schmitt. In this paper, two possible standard error formulas for this statistic are derived and evaluated, hypothesis testing procedures based directly on this descriptive statistic are outlined, and the results of applications to simulated data are presented. Three tables are included. (Contains 29 references.) (Author)

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

ED 382 666

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

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

RR-94-13

Evaluation of the Magnitude of Differential Item Functioning In Polytomous Items

Rebecca Zwick
Dorothy T. Thayer
Educational Testing Service

PROGRAM STATISTICS RESEARCH

Technical Report No. 94-2

Educational Testing Service
Princeton, New Jersey 08541

TM023104

BEST COPY AVAILABLE

The Program Statistics Research Technical Report Series is designed to make the working papers of the Research Statistics Group at Educational Testing Service generally available. The series consists of reports by the members of the Research Statistics Group as well as their external and visiting statistical consultants.

Reproduction of any portion of a Program Statistics Research Technical Report requires the written consent of the author(s).

Evaluation of the Magnitude of Differential Item Functioning In Polytomous Items

Rebecca Zwick
Dorothy T. Thayer
Educational Testing Service

Program Statistics Research
Technical Report No. 94-2

Research Report No. 94-13

Educational Testing Service
Princeton, NewJersey 08541

April 1994

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

Evaluation of the Magnitude of Differential Item Functioning in Polytomous Items

Rebecca Zwick

Dorothy T. Thayer

Educational Testing Service

April 22, 1994

This paper was presented at the annual meeting of the American Educational Research Association, New Orleans, April 7, 1994. We gratefully acknowledge the support of the ETS Research Division, as well as the contributions of John Donoghue, who provided programs from a previous study, John Mazzeo, who reviewed the paper, and Neil Dorans, who provided comments on an earlier version. Reprint requests should be sent to Rebecca Zwick, Educational Testing Service, Princeton, NJ 08541.

Running head: Evaluating DIF in Polytomous Items

Abstract

Several recent studies have investigated the application of statistical inference procedures to the analysis of differential item functioning (DIF) in test items that are scored on an ordinal scale. Mantel's extension of the Mantel-Haenszel test is a possible hypothesis-testing method for this purpose. The development of descriptive statistics for characterizing DIF in polytomous test items has received less attention. A statistic that appears well-suited as a summary index was proposed by Dorans and Schmitt. In this paper, two possible standard error formulas for this statistic are derived and evaluated, hypothesis testing procedures based directly on this descriptive statistic are outlined, and the results of applications to simulated data are presented.

Evaluation of the Magnitude of Differential Item Functioning in Polytomous Items

Several recent studies have investigated the application of statistical inference procedures to the analysis of differential item functioning (DIF) in test items that are scored on an ordinal scale. Several studies (Chang, Mazzeo, & Roussos, 1993; Mazzeo & Chang, 1994; Welch & Hoover, 1993; Zwick, Donoghue, & Grima. 1993a; 1993b) have examined Mantel's (1963) extension of the Mantel-Haenszel (1959) test. An early application of the Mantel approach to DIF data was conducted by Holland and Thayer (Holland, 1991). Other approaches that have been evaluated include combined *t*-tests (Welch & Hoover, 1993), an extension of Shealy and Stout's (1993) SIBTEST procedure (Chang, Mazzeo, & Roussos, 1993; Mazzeo & Chang, 1994), an application of logistic discriminant function analysis (Miller & Spray, 1993), logistic regression approaches (Holland, 1991; Rogers & Swaminathan, 1994), and methods based on item response theory (IRT) (Glas, 1991; Muraki, 1993; Wainer, Sireci, & Thissen, 1991). A review of polytomous DIF methods is given by Potenza and Dorans (in press).

The development of descriptive statistics for characterizing DIF in polytomous test items has received less attention. Dorans and Schmitt (1991) proposed a statistic that appears well-suited for this purpose, but they did not discuss the variability of this index. In the present paper, two possible standard error formulas for the statistic are derived and evaluated, hypothesis testing procedures are described and compared to Mantel's (1963) approach, and analyses of real and simulated test data are presented.

The first section of the paper describes the structure of the data and outlines Mantel's approach. The second section presents the descriptive index proposed by Dorans and Schmitt. The third section gives two alternative ways of deriving a standard error, mentions two additional approaches, and addresses issues of hypothesis testing. The fourth section shows an illustrative analysis, the fifth section presents the results of applying the procedures to simulation data, and the final section lists some issues for future research.

Mantel's Test of Conditional Association for Ordinal Response Variables

Mantel (1963) proposed several extensions to the Mantel-Haenszel (1959) test of conditional association. One of these is a test of whether an ordered response variable is associated with a dichotomous grouping variable, conditional on a third nuisance variable. In the case of DIF analyses, the data are organized into a $2 \times T \times K$ contingency table, where T is the number of response categories and K is the number of levels of a stratification variable, such as score on the entire test. In each of the K strata, the data can be represented as a $2 \times T$ contingency table like that shown in Table 1. "Reference" and "focal" denote the two groups to be compared and the y values, $y_1, y_2, \ldots, y_T$ represent the T possible scores on the item. The body of the table contains values of n_{Rtk} and n_{Ftk} which denote the numbers of reference and focal group members, respectively, who are at the k^{th} level on the stratification variable and received an item score of y_t. A "+" denotes summation over a particular index. For example, n_{F+k} denotes the total number of focal group members in the k^{th} stratum.

Insert Table 1 about here

The statistic proposed by Mantel, reformulated in the notation of this paper and expressed as a Z-statistic rather than a chi-square statistic, is

$$Z = \frac{\sum_k F_k - \sum_k \mathrm{E}(F_k)}{\sqrt{\sum_k \mathrm{Var}(F_k)}} , \qquad (1)$$

where F_k, the sum of scores for the focal group in the k^{th} stratum, is defined as

$$F_k = \sum_t y_t n_{Ftk} . \qquad (2)$$

Treating the row and column marginals within each stratum as fixed, the expectation of F_k under the hypothesis of no association (H_0) is

$$E_H(F_k) = \frac{n_{F+k}}{n_{++k}} \sum_t y_t n_{+tk} \qquad (3)$$

and the variance of F_k under H_0 is

$$Var_H(F_k) = \frac{n_{R+k} n_{F+k}}{n_{++k}^2 (n_{++k} - 1)} \left\{ \left(n_{++k} \sum_t y_t^2 n_{+tk} \right) - \left(\sum_t y_t n_{+tk} \right)^2 \right\} . \qquad (4)$$

Although Mantel does not explicitly state as much, his formulation is consistent with the assumption that, within each stratum, the vector of frequencies for one group (say, the focal group) has a multivariate hypergeometric distribution (Johnson & Kotz, 1969); that is, $\mathbf{n}_{Fk} = (n_{F1k}, n_{F2k}, \ldots, n_{FTk})$ has a multivariate hypergeometric distribution with parameters n_{F+k} and $\mathbf{n}_{+tk}$. (All vectors are of dimension T. The "H" subscript in (3) and (4) denotes the hypergeometric model.) Under H_0, the statistic in (1) is approximately distributed as a standard normal variate. In the case of dichotomous items, this statistic is the same as the Mantel-Haenszel (1959) statistic,[1] which is the basis for the DIF procedure of Holland and Thayer (1988).

In DIF applications, rejection of H_0 suggests that, even after members of the reference and focal groups are matched on some measure of ability (the stratification variable), they tend to differ in their item scores. Using the formulation in (1), a negative Z value implies that, conditional on the stratification variable, the focal group has a lower mean item score than the reference group.

Two previous findings regarding DIF assessment for polytomous items are relevant here. First, Chang and Mazzeo (in press) showed that under certain item response models, including Masters' (1982) partial credit model, identity of conditional item means for two groups implies identity of the set of item category response functions for the groups on that item. (For an item with T response categories, there will be T such response functions.) Therefore, no information is

lost by comparing only the conditional means. Second, Zwick, Donoghue, and Grima (1993a; 1993b) showed that, under the partial credit model, there is a theoretical rationale for defining the stratification variable (say, *S*) as the simple sum of the item scores, including the studied item. Under certain partial credit model assumptions, the odds ratios for any pair of item score categories, conditional on *S*, are constant across values of *S* and are equal to unity when the item response functions for the reference and focal groups are the same. This is an extension to the polytomous case of the findings of Holland and Thayer (1988). This desirable concordance between the IRT-based definition of DIF and the definition based on the odds ratios at each level of *S* can be shown to stem from the fact that *S* is a sufficient statistic for examinee ability in the Rasch and partial credit models (see Lewis, 1993; Masters, 1982; Zwick, 1990).

The *SMD* Index of DIF in Polytomous Items

Dorans and Schmitt (1991) proposed a summary statistic that compares the means of the reference and focal groups, adjusted for differences in the distribution of reference and focal group members across the values of the stratification variable. The statistic can be viewed as an extension of the standardization statistic (*STD P-DIF*) developed by Dorans and Kulick (1986) for summarizing DIF in the case of dichotomous items. In the present paper, the proposed statistic is labeled *SMD* for "standardized mean difference." Reformulated in the notation used here, it is defined as follows:

$$SMD = \sum_k w_{Fk}\, m_{Fk} - \sum_k w_{Fk}\, m_{Rk} \qquad (5)$$

where $w_{Fk} = \frac{n_{F+k}}{n_{F++}}$ is the proportion of focal group members who are at the k^{th} level of the stratification variable, $m_{Fk} = \frac{1}{n_{F+k}} F_k$ is the mean item score for the focal group in the k^{th} stratum, and $m_{Rk} = \frac{1}{n_{R+k}} R_k$ is the analogous value for the reference group, where $R_k = \sum_i y_i\, n_{Rik}$.

As in *STD P-DIF*, the first term of (5) is just the grand mean of the item scores for the focal group. The second term is the mean item score for the reference group, "standardized" as if the reference group distribution across strata was the same as the focal group distribution.

Variability of The *SMD* Index of DIF

Dorans and Schmitt (1991) did not include a standard error for *SMD* and left open the question of "how big is big" (p. 22). As detailed below, several different models can be used to derive a variance formula for *SMD*. First, a general form for Var(*SMD*) will be presented; the result of invoking particular model assumptions will then be explored.

$$\mathrm{Var}(SMD) = \mathrm{Var}\left(\sum_k w_{Fk} m_{Fk} - \sum_k w_{Fk} m_{Rk}\right) = \sum_k w_{Fk}^2 \,\mathrm{Var}\left(\frac{1}{n_{F+k}} F_k - \frac{1}{n_{R+k}} R_k\right)$$

$$= \sum_k w_{Fk}^2 \left\{\left(\frac{1}{n_{F+k}}\right)^2 \mathrm{Var}(F_k) + \left(\frac{1}{n_{R+k}}\right)^2 \mathrm{Var}(R_k) - 2\left(\frac{1}{n_{F+k}}\right)\left(\frac{1}{n_{R+k}}\right)\mathrm{Cov}(F_k, R_k)\right\}. \quad (6)$$

(The weights, $w_{Fk} = \frac{n_{F+k}}{n_{F++}}$, can be treated as fixed quantities here because, in both the models to be discussed, n_{F+k} and n_{F++} are assumed to be fixed.)

Mantel's Multivariate Hypergeometric Model

The multivariate hypergeometric framework of Mantel (1963) can be used to obtain the variance of *SMD* under H_0. Under this model, $\mathrm{Var}(F_k) = \mathrm{Var}_H(F_k)$ is given by (4) and $\mathrm{Var}_H(R_k) = \mathrm{Var}_H(F_k)$; that is, the subscripts R and F are interchangeable in (4). Because the column marginals are fixed, there is a negative covariance between F_k and R_k. Specifically,

$$\mathrm{Cov}_H(F_k, R_k) = \mathrm{Cov}_H\left(\sum_t y_t n_{Ftk}, \sum_t y_t (n_{+tk} - n_{Ftk}) = -\sum_t y_t^2 \,\mathrm{Var}_H(n_{Ftk}) = -\mathrm{Var}_H(F_k)\right)$$

Now, substituting into (6), we have

$$\mathrm{Var}_H(SMD) = \sum_k w_{Fk}^2 \left(\frac{1}{n_{F+k}} + \frac{1}{n_{R+k}}\right)^2 \mathrm{Var}_H\left(F_k\right). \quad (7)$$

This variance formula was first presented in Zwick (1992).

Two-Multinomial Model

Another model that can be used to derive a standard error of *SMD* is the two-multinomial model. Application of this model is a natural extension of the approach used by Phillips and Holland (1987) and by Robins, Breslow, and Greenland (1986) in deriving a standard error for the log of the Mantel-Haenszel odds ratio, which is the basis of the *MH D-DIF* index of DIF. In the two-multinomial model, we assume that, within the k^{th} stratum, the frequencies $\mathbf{n}_{Fk}$ have a multinomial distribution with parameters $\mathbf{\Pi}_{Fk}$ and n_{F+k} and the frequencies $\mathbf{n}_{Rk}$ have a multinomial distribution with parameters Π_{Rk} and n_{F+k}. (All vectors are of dimension T.) This model treats the row, but not the column marginals as fixed and does not invoke the null hypothesis of no association.

Using the properties of the multinomial distribution, we have that

$$\mathrm{Var}_M(F_k) = \mathbf{y}' n_{F+k}(\mathbf{D}_{\Pi Fk} - \mathbf{\Pi}_{Fk}\mathbf{\Pi}'_{Fk})\mathbf{y} \tag{8}$$

and

$$\mathrm{Var}_M(R_k) = \mathbf{y}' n_{R+k}(\mathbf{D}_{\Pi Rk} - \mathbf{\Pi}_{Rk}\mathbf{\Pi}'_{Rk})\mathbf{y}, \tag{9}$$

where $\mathbf{D}_{\Pi Fk}$ is a T x T matrix with the elements of $\mathbf{\Pi}_{Fk}$ on the diagonal and zeroes elsewhere, and $\mathbf{D}_{\Pi Rk}$ is the corresponding matrix for the reference group.

Alternatively, (8) and (9) can be expressed in scalar notation analogous to (4), as follows:

$$\mathrm{Var}_M(F_k) = n_{F+k}\left[\sum_t y_t^2 \pi_{Ftk} - (\sum_t y_t \pi_{Ftk})^2\right] \tag{10}$$

$$\mathrm{Var}_M(R_k) = n_{R+k}\left[\sum_t y_t^2 \pi_{Rtk} - (\sum_t y_t \pi_{Rtk})^2\right]. \tag{11}$$

Now, substituting (10) and (11) into (6) and noting that $\mathrm{Cov}(F_k, R_k)=0$ in this model, we have

$$\mathrm{Var}_{\mathrm{M}}(SMD) = \sum_k w_{\mathrm{F}k}^2 \left\{ \left(\frac{1}{n_{\mathrm{F}+k}} \right)^2 \mathrm{Vai}_{\mathrm{M}}(F_k) + \left(\frac{1}{n_{\mathrm{R}+k}} \right)^2 \mathrm{Var}_{\mathrm{M}}(R_k) \right\} . \qquad (12)$$

In estimating $\mathrm{Var}_{\mathrm{M}}(SMD)$, the elements $\pi_{\mathrm{F}tk}$ are estimated by $\frac{n_{\mathrm{F}tk}}{n_{\mathrm{F}+k}}$ and analogously for $\{\pi_{\mathrm{R}tk}\}$.

Note that multiplying $\hat{\mathrm{Var}}_{\mathrm{M}}(F_k)$ (based on equation 10) by the finite population correction factor $n_{\mathrm{R}+k}/(n_{++k}-1)$ yields $\mathrm{Var}_{\mathrm{H}}(F_k)$ (equation 4). Also, multiplying $\mathrm{Var}_{\mathrm{M}}(F_k)$ by $n_{\mathrm{F}+k}/(n_{\mathrm{F}+k}-1)$, and $\mathrm{Var}_{\mathrm{M}}(R_k)$ by $n_{\mathrm{R}+k}/(n_{\mathrm{R}+k}-1)$ in (12) and substituting the appropriate parameter estimates yields the SIBTEST variance of Shealy and Stout (1993, p. 169, equation 19), provided that weights based on the focal group distribution are used in the Shealy-Stout formula.

Other Models

Two models have been presented here for deriving a variance for *SMD*. Two related models that naturally come to mind are a "one-multinomial" model, in which it is assumed that the multinomial probabilities are the same for the reference and focal groups, and a noncentral multivariate hypergeometric model. The one-multinomial model can be easily derived from equations 10-12 by assuming equality of the within-stratum reference and focal group probabilities, which are then estimated by $\frac{n_{+tk}}{n_{++k}}$. Under this hypothesis of no conditional association,

$$\hat{\mathrm{Var}}_{\mathrm{M}}(SMD) = \sum_k \frac{n_{++k}-1}{n_{++k}} \mathrm{Var}_{\mathrm{H}}(SMD_k) \qquad (13)$$

where

$$SMD_k = w_{\mathrm{F}k}(m_{\mathrm{F}k} - m_{\mathrm{R}k}) \qquad (14)$$

is the k^{th} term of the *SMD* statistic and

$$\mathrm{Var}_{\mathrm{H}}(SMD_k) = w_{\mathrm{F}k}^2 \left(\frac{1}{n_{\mathrm{F}+k}} + \frac{1}{n_{\mathrm{R}+k}} \right)^2 \mathrm{Var}_{\mathrm{H}}(F_k) \qquad (15)$$

is the variance of the k^{th} term of *SMD* under the multivariate hypergeometric model.

Because of the close relationship between the multivariate hypergeometric and multinomial models (Johnson & Kotz, p. 301), the somewhat unwieldy noncentral multivariate hypergeometric model is expected to yield results very similar to (12) in large samples.

Hypothesis Testing

One possible approach to hypothesis testing would be to perform Mantel's test and then use *SMD* as a supplementary descriptive statistic, along with its standard error. However, this approach involves some redundancy because of the close relationship between Mantel's statistic and the statistic *Z(SMD)* formed by dividing *SMD* by its standard error. In terms of (14), the numerator of Mantel's statistic can be expressed as

$$\frac{n_{R+k}\, n_{F+k}}{n_{++k}} \left(m_{Fk} - m_{Rk}\right) = n_{F++} \sum_{k} \frac{n_{R+k}}{n_{++k}} SMD_k \qquad (16)$$

and in terms of (15), the denominator of Mantel's statistic can be written as

$$n_{F+-}^2 \sum_{k} \left(\frac{n_{R+k}}{n_{++k}}\right)^2 \mathrm{Var}_H\left(SMD_k\right) \qquad (17)$$

If the weighting function in (5), rather than in (16), is the desired one, it is intuitively appealing to use *SMD* both as a descriptive measure and as the basis for a test of the hypothesis of no conditional association between group membership and item score. The statistic $Z_H \equiv Z_H(SMD)$ obtained by dividing *SMD* by the square root of (7) and the statistic $Z_M \equiv Z_M(SMD)$ based on the variance in (12) are approximately distributed as standard normal variates under H_0. As noted by Dorans and Kulick (1986), weighting functions other than $w_{Fk} = \frac{n_{F+k}}{n_{F++}}$ may be of interest; the appropriate standard errors can be derived using the same steps presented here.[2]

Application of New Models to *SE(STD P-DIF)*

In the case of a dichotomous item with scores coded 0 and 1, *SMD* reduces to *STD P-DIF*. However, the variance estimates in (7) and (12) are computed under different assumptions from the variance estimate that is used at Educational Testing Service (ETS) for *STD P-DIF*. In the current ETS formulation, the variance is estimated as

$$\hat{\mathrm{Var}}_{\mathrm{ETS}}(STD\,P-DIF) = \mathrm{Var}\left(\sum w_{Fk}\, m_{Fk}\right) + \mathrm{Var}\left(\sum w_{Fk}\, m_{Rk}\right) \quad (18)$$

$$= m_{F+}(1-m_{F+})/n_{F+-} + \sum w_{Fk}^2\, m_{Rk}(1-m_{Rk})/n_{R+k} \quad (19)$$

where m_{Fk} and m_{Rk} are now means of dichotomous variables scored 0 and 1 and $m_{F+} = n_{F1+}/n_{F+-}$. The first parenthesized term in (18) is simply the overall focal group proportion correct, say, m_{F+}, which is treated as a binomial proportion. In computing the variance of the right-hand parenthesized term in (18), the m_{Rk} are treated as binomial proportions within strata and the w_{Fk} are treated as fixed. Based on simulation results, Donoghue, Holland, and Thayer (1993) suggested that "improvements in the estimation of this standard error may be desirable" (p. 165).

A different estimate of $\mathrm{Var}(STD\,P-DIF)$ is obtained by applying (7). In the case of a dichotomous response variable, the term $\mathrm{Var}_H(F_k)$ in (7) can be expressed more simply as

$$\mathrm{Var}_H(F_k) = \frac{n_{R+k}\, n_{F+k}\, n_{+0k}\, n_{+1k}}{n_{++k}^2\,(n_{++k}-1)}, \quad (20)$$

which can be recognized as the within-stratum variance for the Mantel-Haenszel (1959) statistic. These hypergeometric within-stratum variances are equal to binomial within-stratum variances (with the binomial parameter estimated by n_{+1k}/n_{++k}) multiplied by the finite population correction factor, $n_{R+k}/(n_{++k}-1)$.

When $T = 2$, the variance formula in (12) can be expressed as

$$\mathrm{Var}_{\mathrm{M}}(STD\ P-DIF) = \sum_{k} w_{Fk}^{2}\,\pi_{Fk}(1-\pi_{Fk})/n_{F+k} + \sum w_{Fk}^{2}\,\pi_{Rk}(1-\pi_{Rk})/n_{R+k} \quad . \qquad (21)$$

After substituting m_{Rk} as an estimate of π_{Rlk}, the reference group term in (21) is the same as the corresponding term in (19), but substituting m_{Fk} for π_{Flk} does not cause the focal group terms to be the same. It is possible that application of (7) (and (20)) or (21) may provide an estimate of the variance of *STD P-DIF* that is preferable to the one that is currently in use.[3] As a spinoff of the current project, the standard error formulas presented here for *SMD* will be evaluated for the case of dichotomous items.

Hypothetical Example

Suppose that the data for the reference and focal groups are as shown in the top two panels of Table 2 for an item that is scored on a 1-3 scale. Each panel represents one of the two levels of the stratification variable. The entries represent frequencies of examinees; for example, the "5" in the upper left of the top panel indicates that, among low scorers on the stratification variable, five reference group members received an item score of "1."

Insert Table 2 about here

The (unadjusted) difference between the item means for the two groups, obtained by subtracting the reference group mean (2.45) from the focal group mean (2.31), was -0.14. This value, labeled *impact*, is shown at the foot of Table 2, along with other summary statistics. Simply comparing these item means would lead to the conclusion that the focal group performed more poorly on this item than the reference group. Although the *impact* was negative, note that the Mantel Z statistic was positive (0.37), reflecting the fact that *within* each level of the stratification variable, the focal group had a higher item mean than the reference group, as shown in the summary panel of Table 2. The negative *impact* occurred because the reference group members were more likely than the focal group members to receive a high score on the stratification variable

and high scores on the stratification variable were associated with high item scores. The *SMD* adjusts for differences in the distribution of reference and focal groups across levels of the stratification variable by "standardizing" the mean for the reference group and subtracting this standardized mean (2.26) from the focal group mean (2.31). The resulting value of *SMD* was positive (0.05), like the *Z* statistic. This *SMD* value indicates that the conditional between-group difference in mean item score (focal - reference) was 0.05 of a score point, after adjusting for group differences in the distribution of the stratification variable. Using the hypergeometric model that assumes no association (equation 7), the standard error of *SMD* was 0.140; based on the two-multinomial model (equation 12), the standard error was 0.135. The *Z* statistics corresponding to these two models were 0.39 and 0.40, respectively.

Based on the Mantel and *SMD* statistics, the conclusion would be that, after matching examinees on a measure of overall proficiency in the area of interest, the focal group performed better, rather than worse, than the reference group, although the difference was extremely small and not statistically significant. The *SMD* statistic, like the Mantel *Z* statistic, reflects the superior performance of the focal group *within* each level of the stratification variable.

Application to Simulated Data

The standard error in (7) was computed in the simulation study of Zwick, Donoghue and Grima (1993a; 1993b) and compared to the empirical variation of *SMD* across replications. More recently, the same simulation programs were used to generate new data sets so that the standard error formulas in (7) and (12) could be tested and compared. Results appear in Table 3.

Insert Table 3 about here

The top panel of Table 3 provides results for a standard normal focal population; the lower panel corresponds to the case in which the focal population was N(-1, 1). The reference population was standard normal for all simulation conditions.

Item response data were generated for nine studied items, each of which was examined in turn, and for 24 "matching items," used to construct the stratification variable. (No DIF was present in the matching items.) The simulation represented a test consisting of a relatively large set (20) of dichotomous items and a small set (5) of polytomous items. The three-parameter logistic (3PL) model was used to generate the responses for the 20 dichotomous items. (The *b* parameters had a mean of 0 and a standard deviation of 1.37, the *a* parameters had a mean of .88 and a standard deviation of .19, and the *c* parameters were equal to .15.) The partial credit model (Masters, 1982) was used to generate the item responses for the remaining four matching items and for the nine studied items. All partial credit items were assumed to have four response categories; each item, therefore, had three difficulty parameters. (The reference group difficulties, defined as in Masters, 1982, ranged from -2.25 to 3.40. Assuming the inclusion of a scale factor of 1.7, as in the 3PL generating model for the dichotomous items, the *a* parameters for the polytomous items were equal to 1/1.7 = .59.) Note that, although there was reason to expect that the procedures would perform optimally when all items followed the Rasch or partial credit model, the simulation conditions were not completely consistent with this model.

The nine studied items were obtained by crossing three DIF conditions with three sets of reference group item parameters. The three DIF conditions were as follows:

No DIF: The three difficulty parameters for each item were the same for both groups.

Constant (.1): Each difficulty parameter for the focal group was obtained by adding 0.1 to the reference group parameter.

Constant (.25): Each difficulty parameter for the focal group was obtained by adding 0.25 to the reference group parameter.

The parameters for all simulation items appear in Zwick, Donoghue, and Grima (1993b).

For each item within each focal population condition, 100 replications were conducted and the following statistics were computed: Mean and standard deviation across replications of *SMD*, ratio of the mean value of $SE_H(SMD)$ to the standard deviation of *SMD*, ratio of the mean value of $SE_M(SMD)$ to the standard deviation of *SMD*, and proportion of replications in which H_0 was

rejected for the Mantel Z, Z_H, and Z_M statistics, respectively ($\alpha = .05$). Because the variation across items within a DIF condition did not appear to be systematic or meaningful, results were averaged for the three items within each of the three DIF conditions. Results for each focal population condition were analyzed separately.

Several aspects of the results are worthy of note. First, the magnitude of *SMD* was consistent with the generating parameters. In the no-DIF condition, the average *SMD* was close to zero, and in the two constant DIF conditions, the average *SMD* was approximately equal to the item standard deviation in the score point metric (.5), multiplied by the DIF magnitude (0.1 or 0.25), which is expressed in standard deviation units of the ability metric.

The standard error ratios were always smaller for $SE_M(SMD)$ than for $SE_H(SMD)$. This is not surprising: Equation 13 shows that under H_0, $SE_M(SMD)$ must be smaller than $SE_H(SMD)$. In the non-null conditions, the across-replication averages of $SE_M(SMD)$ and $SE_H(SMD)$for the nine items (not shown) never departed by more than .01 from the corresponding averages for the null case.

An aspect of the results that is harder to explain is the relation between the theoretical standard errors and the empirical standard deviation of *SMD*. When the focal population was N(0, 1), both theoretical formulas produced values that tended to be too small in the no-DIF and Constant (.1) condition, but tended to be too large in the Constant (.25) condition. Although $SE_M(SMD)$ does not invoke H_0, its performance in the non-null conditions was not, in general, superior to that of $SE_H(SMD)$.

Averaged over all nine items, the standard error ratios look quite promising in the focal N(0,1) condition: 1.01 for $SE_H(SMD)$ and .97 for $SE_M(SMD)$. In the focal N(-1,1) condition, $SE_H(SMD)$ continued to perform well, with an average ratio of 1.00, but $SE_M(SMD)$ tended to be too small, with an average ratio of .92. In their simulation study of DIF in dichotomous items, Donoghue Holland, and Thayer (1993) computed the ratio of the empirical standard deviation of DIF statistics to the average of the formula-based standard errors (i.e., the reciprocal of the ratios computed here). They reported an average ratio of .96 for *MH D-DIF* and .92 for *STD P-DIF*.

The power of all three Z statistics tended to be smaller in the focal N(-1, 1) condition than in the focal N(0, 1) condition, although Z_H and Z_M both had higher Type I error rates in the focal N(-1, 1) population. It is interesting that in the focal N(-1, 1) condition, the Mantel Z had a smaller Type I error rate than both Z_H and Z_M, but had greater power to detect constant DIF of 0.25. (The standard errors of the tabled rejection rates are approximately 0.01 for the No-DIF condition, 0.02 for the Constant (.1) condition, and 0.03 for the Constant (.25) condition.)

Future Research

The work described here is part of an ongoing research project on DIF for polytomous items. One area that needs to be examined further is the effect of variation in item discrimination on the performance of the Mantel procedure and related methods. Some results of Chang and Mazzeo (1994) show that when the two groups differ in ability and the discrimination parameters of the studied item are very different from those of the matching items, the Type I error rates associated with these methods may be unacceptably high. The extended version of the SIBTEST procedure appears to be more robust to variation in the *a* parameters. (We have conducted some subsequent analyses that show that the degree of inflation for the Mantel approach is much smaller when test length is increased from 25 to 45 items.) In the dichotomous case, Roussos and Stout (1993) conducted a simulation that showed that both the Mantel-Haenszel test and the Shealy-Stout SIBTEST procedure tended to have inflated Type I error rates under certain departures from the Rasch model. Inflation was somewhat more severe for the Mantel-Haenszel when the reference and focal groups had different ability distributions. On the other hand, Chang, Mazzeo, and Roussos (1993), in an earlier simulation study that used the same data-generating programs as those used to produce Table 3, showed that the extended SIBTEST had slightly higher Type I error rates than the Mantel procedure when the groups had different ability distributions.

The superior robustness of SIBTEST under certain conditions is not entirely surprising. The Mantel-Haenszel and Mantel DIF procedures are rendered independent of the group ability

distributions in models for which the matching variable is a sufficient statistic for ability. This occurs when the Rasch or partial credit models hold and the matching variable is number-right score (Zwick, 1990; Zwick, Donoghue & Grima, 1993). The Mantel-Haenszel and Mantel methods show a certain degree of robustness in that they continue to perform well under modest departures from these models. Unlike SIBTEST, however, these methods do not include an explicit correction for measurement error in the matching variable. Any DIF procedure that uses number-right score as a matching variable (or predictor, in the case of logistic regression approaches) is vulnerable to the Type I error inflation that characterizes the Mantel method in some conditions.

More work is needed to determine whether the Mantel procedure and related approaches tend to break down under conditions that are likely to be encountered in practice. In the dichotomous case, Holland (personal communication) has suggested that an index of the degree of departure from the Rasch model would be useful in investigating this issue. In the polytomous case, an index could be developed to measure departures from the partial credit model, for which total score is a sufficient statistic.

Another area in which further research is planned is the development and evaluation of methods for accommodating the sparseness of data that is often evident at some levels of the matching variable. (In particular, problems occur in computing *SMD* when some values of n_{R+k} are zero because m_{Rk} is undefined in that case. Strata with $n_{F+k} = 0$ do not pose an analogous problem since $n_{F+k} = 0$ implies $w_{Fk} = 0$, resulting in elimination of those strata.). In the simulation results reported here, strata in which n_{R+k} was zero were dropped from the analysis, as proposed in Zwick (1992). Shealy and Stout (1993) also exclude strata for which sample sizes are inadequate. As an alternative to exclusion, simple imputation techniques could be used to assign a value to m_{Rk}, as is ordinarily done in computing *STD P-DIF*, or more sophisticated smoothing methods could be applied. Adjustments to the standard error formulas are also needed. Some recent work on smoothed DIF statistics has been conducted by Dorans, Potenza, and Ramsay (1994).

One of the ultimate goals of this research effort is to develop descriptive categories of DIF for polytomous items that are analogous to the A, B, and C classifications used for dichotomous items. As in the dichotomous case, the criteria will involve some combination of statistical significance and effect size. The statistical approaches considered will not be restricted to those presented here, but will include the polytomous extension of SIBTEST and possibly other procedures. Test developers as well as statisticians from testing programs will be consulted and further analyses of simulated and actual test data will be conducted.

Footnotes

[1]The Mantel-Haenszel statistic is often used in the form of a chi-square, rather than a Z-statistic; that is, the statistic is the square of (1), which has a chi-square distribution with one degree of freedom under H_0. In many applications, including DIF analysis, a continuity correction is included; see Holland and Thayer, 1988.

[2]Longford (personal communication, February 9, 1994) suggested a statistical test that resembles Mantel's procedure but uses a slightly different weighting scheme that incorporates within-stratum score variability.

[3]M. Wang (personal communication, October 7, 1992) suggested a formulation of $SE(STD\,P-DIF)$ identical to (20).

[4]Copyright 1993 by Educational Testing Service. All rights reserved.

References

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

Chang, H.-H., Mazzeo, J., & Roussos, L. (April, 1993). *Detecting DIF for polytomously scored items: An adaptation of Shealy-Stout's SIBTEST procedure.* Presented at the annual meeting of the American Educational Research Association and the National Council on Measurement in Education, Atlanta.

Donoghue, J. R., Holland, P. W., & Thayer, D. T. (1993). A Monte Carlo study of factors that affect the Mantel-Haenszel and standardization measures of differential item functioning. In P. W. Holland & H. Wainer (eds.), *Differential Item Functioning*. (pp. 137-166). Hillsdale, NJ: Erlbaum.

Dorans, N. J., & 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. [Also appears in R. E. Bennett & W. C. Ward (eds.), (1993), *Construction Vs. Choice in Cognitive Measurement* (pp. 135-166). Hillsdale, NJ: Erlbaum.]

Dorans, N. J., Potenza, M. T., & Ramsay, J. O. April, 1994). *Smoothed standardization: A small-sample DIF procedure.* Presented at the annual meeting of the American Educational Research Association, New Orleans.

Glas, C. A. W. (1991). *Testing Rasch models for polytomous items: With an example concerning detection of item bias.* Cito Measurement and Research Department Report 91-2. Arnhem, The Netherlands: Cito.

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.

Johnson, N. L., & Kotz, S. (1969). *Discrete distributions.* New York: Wiley.

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.

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. (April, 1994). *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.

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

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

Phillips, A. & Holland, P. W. (1987). Estimation of the variance of the Mantel-Haenszel log-odds-ratio estimate. *Biometrics*, *43*, 425-431.

Potenza, M., & Dorans, N. J. (in press). DIF assessment for polytomously scored items: A framework for classification and evaluation. *Applied Psychological Measurement.*

Robins, J., Breslow, N., & Greenland, S. (1986). Estimators of the Mantel-Haenszel variance consistent in both sparse data and large-strata limiting models. *Biometrics*, *42*, 311-323.

Rogers, H. J., & Swaminathan, H. (April, 1994). *Logistic regression procedures for detecting DIF in nondichotomous item responses.* Presented at the annual meeting of the American Educational Research Association, New Orleans.

Roussos, L. A., & Stout, W. F. (April, 1993). *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., & 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.

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

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

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

Zwick, R., Donoghue, J. 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 No. 93-14). Princeton, NJ: Educational Testing Service.

Zwick, R. (March, 1992). *Application of Mantel's score test to the analysis of differential item functioning in ordinal items.* Technical memorandum Educational Testing Service.

Table 1

Data for the k^{th} Level of the Stratification Variable

	Item Score					
Group	y_1	y_2	y_3		y_T	Total
Reference	n_{R1k}	n_{R2k}	n_{R3k}	...	n_{RTk}	n_{R+k}
Focal	n_{F1k}	n_{F2k}	n_{F3k}	...	n_{FTk}	n_{F+k}
Total	n_{+1k}	n_{+2k}	n_{+3k}	...	n_{+Tk}	n_{++k}

Table 2

Illustrative DIF Analysis of a Polytomous Item

Frequencies of Reference and Focal Group Members Receiving Each Item Score

Low Score on Stratification Variable

	Item Score			
Group	1	2	3	Total
Reference	5	13	7	25
Focal	3	11	6	20
Total	8	24	13	45

High Score on Stratification Variable

	Item Score			
Group	1	2	3	Total
Reference	18	54	108	180
Focal	1	5	9	15
Total	19	59	117	195

Summary Statistics

	Low on Matching Variable		High on Matching Variable		Total	
Statistic	Ref.	Focal	Ref.	Focal	Ref.	Focal
Proportion of cases	0.12	0.57	0.88	0.43	1.00	1.00
Item mean	2.08	2.15	2.50	2.53	2.45	2.31
Standardized mean	-	-	-	-	2.26	-

Impact = 2.31 - 2.45 = -.14

Mantel $Z = 0.37$

$SMD = 2.31 - 2.26 = 0.05$

$SE_H(SMD) = 0.140$, $Z_H = 0.39$

$SE_M(SMD) = 0.135$, $Z_M = 0.40$

Table 3

Results of Nine-Item Simulation

	SMD		$SE_H(SMD)$	$SE_M(SMD)$	P(Reject H_0)[b]		
	Mean	S.D.	Ratio[a]	Ratio	Mantel *Z*	Z_H	Z_M
				Focal N(0, 1)			
DIF Condition							
No DIF	-.00	.05	.99	.96	.05	.04	.05
Constant (.1)	-.05	.05	.95	.92	.18	.17	.20
Constant (.25)	-.12	.04	1.09	1.04	.78	.75	.79
				Focal N(-1, 1)			
No DIF	-.00	.06	1.00	.92	.03	.06	.08
Constant (.1)	-.05	.06	1.00	.92	.18	.15	.21
Constant (.25)	-.13	.06	.99	.92	.73	.62	.69

Note. The reference population was N(0, 1), $n_{R++} = n_{F++} = 500$. Each row in the table corresponds to 3 distinct items. There were 100 replications per item; each table entry is the average result for 3 items.

[a] For each item, the ratio of the across-replication average of SE_H (*SMD*) to the empirical S.D. of *SMD* was computed. The table entry is the average ratio for 3 items. Analogous computations were performed for SE_M (*SMD*).

[b] Each entry is the proportion of rejections for $\alpha = .05$. The standard errors of the rejection rates are approximately 0.01 for the No-DIF condition, 0.02 for the Constant (.1) condition, and 0.03 for the Constant (.25) condition.