Identification of Items that Show Nonuniform DIF

Pankaja Narayanan, Educational Testing Service

H. Swaminathan, University of Massachusetts at Amherst

This study compared three procedures—the Mantel-Haenszel (MH), the simultaneous item bias (SIB), and the logistic regression (LR) procedures—with respect to their Type I error rates and power to detect nonuniform differential item functioning (DIF). Data were simulated to reflect a variety of conditions: The factors manipulated included sample size, ability distribution differences between the focal and the reference groups, proportion of DIF items in the test, DIF effect sizes, and type of item. 384 conditions were studied. Both the SIB and LR procedures were equally powerful in detecting nonuniform DIF under most conditions. The MH procedure was not very effective in identifying nonuniform DIF items that had disordinal interactions. The Type I error rates were within the expected limits for the MH procedure and were higher than expected for the SIB and LR procedures; the SIB results showed an overall increase of approximately 1% over the LR results. *Index terms: differential item functioning, logistic regression statistic, Mantel-Haenszel statistic, nondirectional DIF, simultaneous item bias statistic, SIBTEST, Type I error rate, unidirectional DIF.*

In recent years, there has been concern over the issue of differential item functioning (DIF) in educational data. DIF is said to exist if examinees having the same underlying ability have different probabilities of getting an item correct regardless of group membership. From an item response theory (IRT) perspective, an item shows DIF if the item response functions (IRFs) evaluated across two different subgroups are not identical.

According to Mellenbergh (1982), two types of DIF can occur in educational dichotomous data. Uniform DIF occurs when there is no interaction between ability level and group membership. Nonuniform DIF occurs when there is interaction between ability level and group membership. In general, although uniform DIF occurs more often than nonuniform DIF in standardized tests, nonuniformly functioning items have been identified in real data (Hambleton & Rogers, 1989; Linn, Levine, Hastings, & Wardrop, 1981; Mellenbergh, 1982).

A variety of statistical procedures have been developed for detecting DIF (Berk, 1982; Millsap & Everson, 1993). IRT provides a general framework for studying DIF. Unfortunately, IRT-based DIF procedures require large sample sizes, a condition that is often difficult to meet in practice. Because of this, researchers have developed parametric and nonparametric methods to identify DIF that are effective and, at the same time, easy to implement in practice.

Some of the most promising nonparametric methods for detecting DIF are the Mantel-Haenszel (MH) procedure (Holland & Thayer, 1988), the standardization procedure (Dorans & Kulick, 1986), and the simultaneous item bias procedure (SIBTEST, henceforth referred to as SIB; Shealy & Stout, 1993). MH and SIB share a common framework. They are computationally simple, inexpensive, easy to implement in practice, and do not require large sample sizes. Also, both procedures provide statistics that have associated tests of significance.

Swaminathan & Rogers (1990) presented a logistic regression (LR) procedure and demonstrated that it can be implemented easily in practice. A major advantage of the LR procedure is that it is a model-based procedure with the ability variable treated as continuous. It also allows for testing the hypothesis of no interaction

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 20, No. 3, September 1996, pp. 257–274

0146-6216/96/030257-18$2.15

between the ability variable and the group variable. In fact, the MH procedure can be conceptualized as being based on the LR model in which the ability variable is treated as discrete and no interaction between the ability variable and group membership is permitted. The LR procedure is therefore expected to improve on the MH procedure for detecting nonuniform DIF. Previous research has shown that the MH, SIB, and LR procedures are equally effective in the identification of uniform DIF (Ackerman, 1992; Narayanan & Swaminathan, 1994; Rogers & Swaminathan, 1993a; Roussos & Stout, 1993; Swaminathan & Rogers, 1990).

Swaminathan & Rogers (1990) distinguished two types of nonuniform DIF. Ability typically falls in the range −3 to +3 on the ability level scale in item response theory. When the IRFs cross in the middle of this range, a type of nonuniform DIF occurs that is analogous to a disordinal interaction in analysis of variance (ANOVA) models (Swaminathan & Rogers, 1990). When the IRFs cross outside this range or when the IRFs are not parallel but do not cross (a situation that may occur with the three-parameter IRT model), a type of nonuniform DIF analogous to ordinal interaction occurs. Li & Stout (1993) termed these two types of DIF "nondirectional" and "unidirectional," respectively. Using simulated data, Rogers & Swaminathan (1993a) showed that although the MH procedure is capable of detecting ordinal or unidirectional DIF, it is not capable of identifying disordinal or nondirectional DIF. The LR procedure can adequately identify nondirectional DIF because it includes a term for interaction between group membership and ability. The major advantage of the LR procedure is that it can be expanded to condition on more than one ability variable.

Recently, a modification of the SIB procedure, known as crossing-SIBTEST (CRO-SIB), was developed (Li & Stout, 1993). CRO-SIB is designed to detect nonuniform DIF and has the potential for conditioning on more than one ability variable. However, the CRO-SIB procedure has not been studied extensively. Also, to date, there have not been any studies comparing CRO-SIB with MH and LR. Given the possibility that it could be superior to the MH and LR procedures in some situations, a detailed investigation of the three procedures is important and timely.

The main objectives of this study were: (1) to investigate and compare the Type I error rates and power of the MH, CRO-SIB, and LR procedures, and (2) to determine the conditions under which each procedure is optimal for detecting nonuniform DIF.

Description of the DIF Statistics

The Mantel-Haenszel Procedure

The MH procedure (Holland & Thayer, 1988) compares the probabilities of a correct response in the focal and reference groups for examinees of the same ability as reflected in total number-correct score. The group an item is suspected of favoring is referred to as the *reference group*; the group in which an item is suspected of differentially functioning is called the *focal group*. In order to compare the probabilities of a correct response, item response data for the reference and the focal group members are arranged into a series of 2×2 contingency tables. One table is constructed for each test item to accommodate group × item response at each score level. In all, K 2×2 contingency tables are constructed, where K is the number of unique scores for the test. The 2×2 contingency table for the ith item and jth score level is shown in Table 1.

The null DIF hypothesis is that the odds of correctly answering the item at a given score level j are the

Table 1
2×2 Contingency Table at the jth Score Level

	Score on Studied Item		
Group	1	0	Total
Reference	A_j	B_j	N_{Rj}
Focal	C_j	D_j	N_{Fj}
Total	N_{1j}	N_{0j}	$N_{.j}$

same for the reference and the focal group at all K levels of the matching variable. The null and alternate constant odds ratio hypothesis at score level j can be expressed as

$$H_0: \left[\pi_{Rj}/\left(1-\pi_{Rj}\right)\right] = \left[\pi_{Fj}/\left(1-\pi_{Fj}\right)\right] \quad j = 1,2,\ldots,k, \tag{1}$$

and

$$H_A: \left[\pi_{Rj}/\left(1-\pi_{Rj}\right)\right] = \alpha\left[\pi_{Fj}/\left(1-\pi_{Fj}\right)\right] \quad j = 1,2,\ldots,k, \quad \alpha \neq 1, \tag{2}$$

where π_{Rj} is the probability that a reference group (R) examinee with total score j will answer the studied item correctly, and π_{Fj} is the probability that a focal group (F) examinee with total score j will provide a correct answer to the studied item.

Equations 1 and 2 presume uniform DIF, if DIF exists. Uniform DIF is said to occur when the difference in the probability of a correct answer to an item between two groups is constant across all ability levels. The parameter α is called the common odds ratio. When the value of α is equal to 1.0, the probability of a correct response is equal for both groups. A value of α greater than 1.0 indicates that reference group members are more likely to answer the item correctly. Similarly, a value of α less than 1.0 indicates that focal group members are more likely to answer the item correctly. An estimate of the common odds ratio α, known as α_{MH}, also provides an estimate of DIF effect size. It can be expressed as

$$\alpha_{MH} = \frac{\Sigma A_j D_j / N_{..j}}{\Sigma B_j C_j / N_{..j}}. \tag{3}$$

From the K 2×2 tables for a given item, the MH statistic, χ^2_{MH}, with a continuity correction is computed as

$$\chi^2_{MH} = \frac{\left[\left|\Sigma A_j - \Sigma E\left(A_j\right)\right| - .5\right]^2}{\Sigma \mathrm{Var}\left(A_j\right)}, \tag{4}$$

where A_j is the observed number of examinees in the reference group at score level j answering the item correctly,

$$E\left(A_j\right) = \frac{N_{Rj} N_{1\cdot j}}{N_{..j}}, \tag{5}$$

and

$$\mathrm{Var}\left(A_j\right) = \frac{N_{Rj} N_{Fj} N_{1\cdot j} N_{0\cdot j}}{\left(N_{..j}\right)^2 \left(N_{..j} - 1\right)}. \tag{6}$$

The Simultaneous Item Bias Procedure

The SIB procedure (Shealy & Stout, 1993) emphasizes the examination of DIF at the test level and provides a statistical test to detect if DIF is present in one or more items on a test simultaneously. To test whether a set of items in the test is functioning differentially, item response data for the reference and focal groups are formed into two subtests—a "suspect" subtest containing the items that are to be tested for DIF (this can be one or more items) and a "valid" subtest containing the items that measure the construct that the test is purported to measure (i.e., those items not suspected of functioning differentially). To calculate the SIB statistic, examinee subtest scores on the valid subtest are used to group the reference and focal groups into score levels so that, for n items in the test, the number of score levels on the valid subtest will be equal to (at most) $n + 1$. Then, for reference and focal group members with the same valid subtest

scores, the average proportion correct (across examinees) on the suspect subtest is calculated.

Shealy & Stout's (1991) DIF index, β_U, is a parameter denoting the amount of unidirectional DIF (the noncrossing type of DIF in which the same group has a higher proportion correct at all valid subtest score levels). A β_U value of .1 indicates that the average difference in the probabilities of correct response of the "studied" subtest score between reference and focal group examinees at the same ability level is .1.

For unidirectional DIF, the hypothesis of interest is

$$H_0: \beta_U = 0 \text{ vs. } H_A: |\beta_U| > 0. \tag{7}$$

For nondirectional DIF (the modification of SIB for detecting nonuniform DIF called CRO-SIB), the hypothesis of interest is

$$H_0: \beta_C = 0 \text{ vs. } H_A: |\beta_C| > 0. \tag{8}$$

The two hypotheses are tested simultaneously. To control for Type I error, the Bonferroni adjustment is used; that is, each hypothesis is tested at the $\alpha/2$ level of significance so that the overall Type I error rate does not exceed α.

Let

$$X = \sum_{i=1}^{n} U_i \tag{9}$$

be the total score on the valid subtest, where U_i denotes the response to item i scored as 0 or 1, and

$$Y = \sum_{i=n+1}^{N} U_i \tag{10}$$

be the total score on the studied subtest. Let $\bar{Y}_{Rk}$ and $\bar{Y}_{Fk}$ be the average score in the suspect subtest for all examinees in the reference and the focal groups, respectively, attaining a valid subtest score $X = k$, ($k = 0, 1, 2, \ldots, n$). Because $(\bar{Y}_{Rk} - \bar{Y}_{Fk})$ is the difference in performance in the suspect subtest across the two groups among examinees of the same ability, it will equal 0.0 if the suspect subtest items do not show DIF. However, when there are differences in the ability distribution of the reference and the focal groups, even in the case of no DIF, $(\bar{Y}_{Rk} - \bar{Y}_{Fk})$ will differ systematically from 0.0 and will tend to indicate the presence of DIF even though no DIF is present (Shealy & Stout, 1993). Therefore, if differences in ability distributions of the reference and focal groups exist, a model-based adjustment known as the regression correction is used on the means of $\bar{Y}_{Rk}$ and $\bar{Y}_{Fk}$.

According to Shealy & Stout (1993), with the regression correction in place, cautions about the observed score as the matching criterion in place of true scores do not apply to the SIB procedure. [For more details on the classical test theory and item response theory-based justification for the regression correction, see Shealy & Stout (1993).] It follows that an estimate $\hat{\beta}_U$ of β_U is

$$\hat{\beta}_U = \sum_{k=0}^{n} \hat{p}_k (\bar{Y}_{Rk} - \bar{Y}_{Fk}), \tag{11}$$

where $\hat{p}_k$ is the proportion among the focal group examinees attaining a score of $X = k$ on the valid subtest.

The SIB test statistic, B_U, for testing the hypothesis of no uniform DIF is

$$B_U = \hat{\beta}_U / \hat{\sigma}(\hat{\beta}_U), \tag{12}$$

where $\hat{\sigma}(\hat{\beta}_U)$ is the estimated standard error of β_U. The expression for $\hat{\sigma}(\hat{\beta}_U)$ is given in Shealy & Stout (1993).

An estimate $\hat{\beta}_C$ of β_C is defined as

$$\hat{\beta}_C = \sum_{k=0}^{k_0-1} \hat{p}_k\left(\bar{Y}_{Fk} - \bar{Y}_{Rk}\right) + \sum_{k=k_0+1}^{n} \hat{p}_k\left(\bar{Y}_{Rk} - \bar{Y}_{Fk}\right), \tag{13}$$

where k_o is the valid subtest score at which crossing is estimated to occur. A detailed description of how the crossing point is estimated is given in Li & Stout (1993, p. 3). The SIB test statistic B_C for testing the hypothesis of nonuniform DIF is defined as

$$B_C = \hat{\beta}_C / \hat{\sigma}\left(\hat{\beta}_C\right). \tag{14}$$

The expression for $\hat{\sigma}(\hat{\beta}_C)$ is given in Li & Stout (1993). The SIB statistics B_U and B_C have approximate N(0,1) distributions when no DIF is present. The null hypothesis of no DIF is rejected if the value of B_U or B_C exceeds the upper $100(1 - \alpha)$th percentile point of the standard normal distribution.

The Logistic Regression Procedure

The standard equation of the LR model for predicting the dichotomous response variables given a set of independent variables (Bock, 1975) is

$$P\left(u_{pj} = 1 \middle| \theta_{pj}\right) = \frac{\exp\left(\beta_{0j} + \beta_{1j}\theta_{pj}\right)}{1 + \exp\left(\beta_{0j} + \beta_{1j}\theta_{pj}\right)} \quad p = 1, \ldots, n_j,\ j = 1, 2, \tag{15}$$

where

$P(u_{pj} = 1)$ is the response of person p in group j to the item,
β_{0j} is the intercept parameter,
β_{1j} is the slope parameter for group j,
θ_{pj} is the observed ability of person p in group j.

According to the definition of DIF, an item is unbiased if individuals having the same ability have different probabilities of answering an item correctly. Therefore, in the above model, if $\beta_{01} = \beta_{02}$ and $\beta_{11} = \beta_{12}$, it follows that the LR functions for the two groups are the same and the item is unbiased.

By definition, an item exhibits uniform DIF if the LR functions for the two groups are parallel but have different intercepts (i.e., if $\beta_{01} \neq \beta_{02}$, but $\beta_{11} = \beta_{12}$). An item exhibits nonuniform DIF if there is an interaction between the ability level and group membership (i.e., $\beta_{11} \neq \beta_{12}$). In the case of nonuniform DIF, the LR functions are not parallel.

The LR model can be reparameterized to include a parameter corresponding to uniform DIF and a parameter corresponding to nonuniform DIF in the form

$$P\left(u_{pj} = 1\right) = \frac{\exp\left(z_{pj}\right)}{1 + \exp\left(z_{pj}\right)}, \tag{16}$$

where

$$z_{pj} = \tau_0 + \tau_1\theta_{pj} + \tau_2 g_j + \tau_3\left(\theta_{pj} g_j\right), \tag{17}$$

and where

$P(u_{pj})$ is the probability of a correct response for person p in group j,
τ_0 is the intercept,
τ_1 is the coefficient of ability,
τ_2 $(= \beta_{01} - \beta_{02})$ is the group difference,
τ_3 $(= \beta_{11} - \beta_{12})$ is the interaction between group and ability, and

g represents group membership, so that

$$g_p = \begin{cases} .5 \text{ if } j = 1 \\ -.5 \text{ if } j = 2 \end{cases}. \tag{18}$$

In the above model, an item exhibits uniform DIF if $\tau_2 \neq 0$ and $\tau_3 = 0$, and nonuniform DIF if $\tau_3 \neq 0$ (whether or not $\tau_2 = 0$). In the model given by Equation 17, the parameters of each item can be estimated using maximum likelihood estimation. Estimation proceeds by maximizing the likelihood function given by

$$L(u_{pj}|\theta) = \prod_{p=1}^{N}\prod_{j=1}^{n} P(u_{pj})^{u_{pj}}\left[1 - P(u_{pj})\right]^{1-u_{pj}}, \tag{19}$$

where

N is the sample size,
n is the test length,
$u_{pj} = 1$, and
$P(u_{pj})$ is as defined above in Equation 16.

The estimates of the parameters obtained by the maximum likelihood procedure are asymptotically multivariate normally distributed with mean vector $\boldsymbol{\tau}$ (the true values of the parameters) and variance-covariance matrix $\boldsymbol{\Sigma}$, the inverse of which is equal to the negative of the expected value of the matrix of second derivatives of the log-likelihood function. In this case, the expected value of the matrix of second derivatives is equal to the matrix of second derivatives (Swaminathan & Rogers, 1990). Thus,

$$\hat{\boldsymbol{\tau}} \sim N(\boldsymbol{\tau}, \boldsymbol{\Sigma}), \tag{20}$$

where $\hat{\boldsymbol{\tau}}' = [\tau_0\ \tau_1\ \tau_2\ \tau_3]$, and $\hat{\boldsymbol{\tau}}$ is its estimate. The asymptotic standard error of the estimate of τ_s ($s = 0, 1, 2, 3$) is the square root of the sth diagonal element of $\boldsymbol{\Sigma}$; that is,

$$SE(\hat{\tau}_s) = \left[\boldsymbol{\Sigma}_{ss}\right]^{1/2}. \tag{21}$$

Testing hypotheses regarding the presence of DIF in test items requires testing hypotheses about some of the elements of $\boldsymbol{\tau}$. The hypotheses of interest are H_0: $\tau_2 = 0$ and H_0: $\tau_3 = 0$. Because the estimates of τ_2 and τ_3 are univariate normal, these hypotheses can be tested individually, and the overall Type I error rate is controlled using a Bonferroni adjustment as with the SIB procedure. Alternatively, the two hypotheses $\tau_2 = 0$ and $\tau_3 = 0$ can be tested simultaneously as

$$H_0: \mathbf{C}\boldsymbol{\tau} = \mathbf{0} \tag{22}$$

against

$$H_0: \mathbf{C}\boldsymbol{\tau} \neq \mathbf{0}, \tag{23}$$

where $\mathbf{C}$ is a (2×4) matrix defined as:

$$\mathbf{C} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{24}$$

The statistic for testing the joint (multivariate) hypothesis given by Equations 22–23 is:

$$\chi^2 = \hat{\boldsymbol{\tau}}'\mathbf{C}'(\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}')^{-1}\mathbf{C}\hat{\boldsymbol{\tau}}. \tag{25}$$

Equation 25 has a χ^2 distribution with two degrees of freedom. When the test statistic exceeds $\chi^2_{\alpha;2}$, the hypothesis of no DIF is rejected. The test in Equation 25 is a multivariate test of the hypothesis given by Equations 22–23.

Power Study for MH, SIB, and LR When DIF Is Present

Method

Simulated datasets were used in this study. Only with simulated datasets was it possible to accommodate prespecified levels of a variety of factors for the MH, CRO-SIB, and LR procedures and to determine the detection rates accurately under the desired conditions.

Factors manipulated. A number of factors that can have an impact on DIF detection rates have been identified in previous research (Mazor, Clauser, & Hambleton, 1992; Rogers & Swaminathan, 1993a; Swaminathan & Rogers, 1990). One main factor of interest is sample size (N). Previous research has shown that the power of MH, CRO-SIB, and LR increased as N increased. However, in practice, the sample size of the focal group is likely to be small. Therefore, in this study, focal group sample sizes of $N = 500$ examinees or less were used.

Test length can have an impact on DIF detection rates because a longer test is likely to produce more reliable scores and, hence, more accurate ability estimates. However, increasing the proportion of items showing DIF is likely to produce a contaminated conditioning variable and this may affect the detection rates. Hence, the power of the DIF procedures is likely to increase when the test length increases and decrease when the proportion of DIF items in the test increases.

Mazor et al. (1992) investigated the impact on MH with two subgroups sampled from equal and unequal ability distributions. Differences in the ability distributions of the groups being compared will have an effect on DIF detection (Mazor et al.; Shealy & Stout, 1993). DIF effect size or the amount of DIF contained in an item is another factor that is likely to have an effect on the DIF detection procedures. As DIF effect size increases, the detection rates of the procedures are expected to increase as well.

In this study, five factors were manipulated: sample size, proportion of items containing DIF, ability distribution differences, DIF effect size, and type of item (combination of difficulty and discrimination). The two reference group sample sizes ($N_R = 500$ and $N_R = 1{,}000$) were crossed with the two focal group sample sizes ($N_F = 200$ and $N_F = 500$) to produce four conditions related to sample size. Test length was not manipulated. Standardized achievement and ability tests usually range from approximately 35 items to 80 items. The study was confined to a single test length of 40 items to investigate the capability of the procedures to detect DIF in a "short" test.

The impact of differences in underlying ability distributions was investigated by examining two different conditions. In the first case, ability distributions for the two groups were set to be equal with mean 0.0 and standard deviation (SD) of 1. In the second case, the mean was set to 0.0 and −1.0 for the reference and the focal groups, respectively, again with both SDs set to 1. Distributions that differed by 1 SD were selected to simulate cases sometimes found in DIF studies (e.g., Hambleton & Rogers, 1989). Because the proportion of DIF items can contaminate the conditioning variable, the proportion of items containing DIF was set at three levels: 0%, 10%, and 20%.

DIF effect size was manipulated using the area between the IRFs for the two groups as the measure of DIF effect size. The area between the IRFs for the two groups can be computed using the formula given by Raju (1988). Four levels of DIF effect size corresponding to area values of .4, .6, .8, and 1.0 were selected to reflect DIF effect size values ranging from a small amount of DIF to a fairly large amount of DIF.

Nonuniform DIF was simulated by keeping the difficulty parameters (bs) for the two groups the same and varying the discrimination parameters (as) for the two groups (see Table 2). 16 items showing nonuniform DIF were simulated by varying the level of the common b—low ($b = -1.5$), medium ($b = 0.0$), and high ($b = 1.5$); the level of a for the two groups—low ($.40 \le a \le .50$ in the reference group and $.72 \le a \le 1.03$ in the focal group) and high ($.47 \le a \le .90$ in the reference group and $1.68 \le a \le 2.01$ in the focal group); and DIF effect size (area values of .4, .6, .8, and 1.0). Four item types were studied: (1) low b, high a; (2) medium b, low a;

(3) medium *b*, high *a*; and (4) high *b*, low *a*.

To simulate a 40-item test with 10% of the items showing DIF (i.e., four items), and to accommodate the characteristics of items that may affect DIF detection, it was necessary to distribute the 16 items into four 40-item tests. Similarly, to simulate 20% of the items showing DIF (i.e., eight items), the 16 DIF items were distributed into two 40-item tests. Item parameter values for the non-DIF items remained the same in all 40-item tests. They were randomly selected from published item parameter values from an administration of the Graduate Management Admission Test (Kingston, Leary, & Wightman, 1988). The *c* parameters for all items were set equal to .20.

Simulation procedures. Data for the study were simulated according to the three-parameter logistic model using the program DATAGEN (Hambleton & Rovinelli, 1973) in order to determine the viability of the three methods in identifying the 16 nonuniform DIF items described above. Nonuniform DIF was simulated by selecting different *a* parameters for the two groups and keeping the *b*s the same for the two groups. The DIF statistics values for MH and LR were obtained using the program DICHODIF (Rogers & Swaminathan, 1993b). The CRO-SIB statistics values were obtained using the program CSIBTEST (Li & Stout, 1994a, 1994b). The item parameter values for the 16 nonuniform DIF items are shown in Table 2 and those for the non-DIF items are shown in Table 3.

Table 2
Item Parameters Used to Generate Nonuniform DIF Items

Item Type and Item Number	DIF Effect Size	a_R	b_R	a_F	b_F
Low *b*, High *a*					
1	.4	.90	−1.50	2.01	−1.50
2	.6	.70	−1.50	1.97	−1.50
3	.8	.56	−1.50	1.79	−1.50
4	1.0	.47	−1.50	1.68	−1.50
Medium *b*, Low *a*					
5	.4	.50	0.00	.72	0.00
6	.6	.46	0.00	.80	0.00
7	.8	.43	0.00	.91	0.00
8	1.0	.40	0.00	1.03	0.00
Medium *b*, High *a*					
9	.4	.90	0.00	2.01	0.00
10	.6	.70	0.00	1.97	0.00
11	.8	.56	0.00	1.79	0.00
12	1.0	.47	0.00	1.68	0.00
High *b*, Low *a*					
13	.4	.50	1.50	.72	1.50
14	.6	.46	1.50	.80	1.50
15	.8	.43	1.50	.91	1.50
16	1.0	.40	1.50	1.03	1.50

Thus, DIF analyses were implemented with datasets simulated for four combinations of sample size, two levels of ability distribution differences, three levels of percent of items containing DIF, four levels of DIF effect size, and four types of items (combinations of *a* and *b*). 384 conditions were studied to investigate nonuniform DIF. The data were replicated 100 times for each condition. The power and Type I error rates of the three statistics were evaluated at $\alpha = .05$ and $\alpha = .01$.

Computation of DIF. In computing the MH DIF statistics, a two-stage procedure recommended by Holland & Thayer (1988) was used. In the first stage, the total score based on all items was used as the

Table 3
Item Parameters for the Non-DIF Items

Item Number	*a*	*b*	*c*	Item Number	*a*	*b*	*c*
1	.44	−.30	.20	21	.92	1.13	.20
2	.55	−1.06	.20	22	.64	−1.55	.20
3	.82	1.02	.20	23	1.01	.81	.20
4	.52	−1.96	.20	24	.61	−.53	.20
5	1.02	1.28	.20	25	.70	1.05	.20
6	.82	.61	.20	26	1.02	.64	.20
7	.92	.42	.20	27	.48	2.12	.20
8	.65	1.68	.20	28	1.01	.91	.20
9	.56	−2.70	.20	29	.53	.87	.20
10	.29	−1.39	.20	30	.36	−2.63	.20
11	.35	−1.12	.20	31	1.12	−1.21	.20
12	.31	−1.37	.20	32	.86	−.57	.20
13	1.05	.10	.20	33	.59	−1.29	.20
14	.51	−.09	.20	34	.56	.40	.20
15	.73	.61	.20	35	1.09	1.11	.20
16	.88	.95	.20	36	.88	−.93	.20
17	1.11	−.35	.20	37	.96	−1.21	.20
18	1.32	.57	.20	38	1.06	2.11	.20
19	.55	1.09	.20	39	.92	.62	.20
20	1.40	1.64	.20	40	.75	−1.01	.20

Note. Item parameters for Items 1–36 did not vary across conditions.

matching criterion to group the examinees, and items showing DIF were identified. In the second stage, items showing DIF (with the exception of the studied item) were excluded from the calculation of the total score used to group examinees. The two-stage procedure described above was not used for the CRO-SIB and LR DIF statistics.

Analysis. A completely crossed five-way ANOVA was used to determine the effects of the five factors on the performance of MH, CRO-SIB, and LR. The dependent variable was the mean detection rate for each of the three procedures. The independent variables were the five factors manipulated in the study.

Results

The ANOVA results presented in Table 4 show that for all three procedures, three of the five factors—*N*, type of item, and DIF effect size—had significant main effects at $\alpha = .05$. In addition, several two-way interaction effects observed were common for the three procedures. These were $N \times$ ability distribution, type of item × ability distribution, ability distribution × DIF effect size, proportion of DIF (% DIF) × DIF effect size, and type of item × DIF effect size. For MH, there were interaction effects for $N \times$ type of item and % DIF × type of item. Table 5 shows the mean detection rates and mean Type I error rates for each of the five factors.

Effect of sample size. The results in Table 5 indicate that the detection rates for the three procedures showed a steady increase as *N* increased. In particular, the detection rates for the three procedures appeared to increase more for an increase in N_F than for an increase in N_R. For example, at $\alpha = .05$, for $N_R = 500$, when N_F increased from 200 to 500, the increase in the detection rates was 10% for MH, 14% for CRO-SIB, and 17% for LR. For $N_R = 1{,}000$, the increase was 14% for MH, 17% for CRO-SIB, and 18% for LR. However, when N_R increased from 500 to 1,000, the increase in the detection rates for the three procedures was 4% to 5% for $N_F = 200$, and approximately 6% to 8% for $N_F = 500$.

For $\alpha = .05$, CRO-SIB showed an increase of approximately 5% in detection rates over LR for all *N*s. The

Table 4
Main Effects and Two-Way Interaction Effects From the ANOVA of the Effects of All Factors on the Performance of the MH, CRO-SIB, and LR Procedures in Detecting DIF

	MH		CRO-SIB		LR	
Factor	*F*	*p*	*F*	*p*	*F*	*p*
Main Effects						
N	29.46	0.00*	60.43	0.00*	46.34	0.00*
Ability Distribution	1.41	.24	4.43	.01*	74.44	0.00*
% DIF	0.00	.97	.02	.89	6.14	.01*
Type of Item	281.98	0.00*	215.94	0.00*	187.27	0.00*
DIF Effect Size	68.75	0.00*	193.23	0.00*	140.56	0.00*
Interaction Effects						
N × Ability Distribution	4.78	0.00*	.69	0.00*	.89	0.00*
N × % DIF	.02	.99	.44	.73	.06	.98
N × Type of Item	4.30	0.00*	1.66	.11	.97	.47
N × DIF Effect Size	.40	.93	.87	.56	.99	.45
Ability Distribution × % DIF	.46	.50	1.40	.24	3.20	.08
Ability Distribution × Type of Item	164.95	0.00*	9.61	0.00*	9.28	0.00*
Ability Distribution × DIF Effect Size	4.64	0.00*	4.80	0.00*	2.53	.05*
% DIF × Type of Item	3.34	.02*	.87	.46	.72	.54
% DIF × DIF Effect Size	4.01	0.00*	9.10	0.00*	5.20	0.00*
Type of Item × DIF Effect Size	12.28	0.00*	9.92	0.00*	4.79	0.00*

*Significant at $p < .05$.

detection rates for MH varied from approximately 31% to approximately 49% for the four *N*s.

The Type I error rates presented in Table 5 show that they were within the nominal limits for MH for all *N*s. They were higher than expected for CRO-SIB and LR, with CRO-SIB showing an increase of approximately .3% over LR. For all three procedures, the Type I error rates were slightly less for the smallest *N* than for other *N*s. For example, for $N_R = 500$ and $N_F = 200$ at $\alpha = .05$, the Type I error rates were approximately 4.5% for MH, 7.8% for CRO-SIB, and 7.5% for LR. For $N_R = 1{,}000$ and $N_F = 500$, the Type I error rates increased to 5.6% for MH, 9.1% for CRO-SIB, and 8.9% for LR.

Effect of ability distribution differences. For all three procedures, the detection rates were higher when examinees were sampled from the equal ability distribution than from the unequal ability distribution (see Table 5). Although the differences in detection rates for the two types of distributions were only approximately 2% to 3% for MH and CRO-SIB, they were much higher (approximately 14%) for LR. For example, at $\alpha = .05$, for the unequal ability distribution, the detection rates decreased from those of the equal ability distribution from 40% to 38% for MH, from 69% to 66% for CRO-SIB, and from 70% to 56% for LR.

For all three procedures, Table 5 shows that the Type I error rates were higher for the unequal ability distribution than those for the equal ability distribution. For example, at $\alpha = .05$, the Type I error rates for the equal ability distributions were 4.1% for MH, 7% for CRO-SIB, and 6.1% for LR. For the unequal ability distributions, they were 5.5% for MH, 10% for CRO-SIB, and 9.8% for LR. Although the increase was only marginal for MH (approximately 1.5%), it was approximately 3% for CRO-SIB and approximately 3.5% for LR. Overall, CRO-SIB showed the highest Type I error rates under both conditions.

Effect of percent of items containing DIF. The detection rates for MH and CRO-SIB did not differ much whether the tests had 10% or 20% DIF items. For LR, there was an increase of approximately 4% for tests with 10% DIF items over tests with 20% DIF items. For example, at $\alpha = .05$, the detection rates were 39% for MH and 68% for CRO-SIB for tests with 10% or 20% DIF items. For LR, the detection rate was 65% when tests contained 10% DIF items and 61% when tests contained 20% DIF items.

The Type I error rates were within nominal limits for MH whether tests contained 0%, 10%, or 20% DIF

Table 5
Mean Percent Detection Rates (Power) and Type I Error Rates for the MH, CRO-SIB, and LR Procedures Under All Conditions for $\alpha = .05$ and $\alpha = .01$

	MH		CRO-SIB		LR	
Factor	.05	.01	.05	.01	.05	.01
Mean Percent Detection Rates (Power)						
Sample Size						
$N_R = 500$, $N_F = 200$	31	18	58	41	52	33
$N_R = 500$, $N_F = 500$	41	29	72	57	69	55
$N_R = 1{,}000$, $N_F = 200$	35	22	62	47	57	37
$N_R = 1{,}000$, $N_F = 500$	49	37	79	68	75	63
Ability Distribution						
Equal	40	31	69	56	70	54
Unequal	38	22	66	51	56	40
% DIF						
10%	39	26	68	53	65	49
20%	39	27	68	53	61	45
Type of Item						
Low *b*, High *a*	66	54	88	78	90	77
Medium *b*, Low *a*	15	6	47	30	44	26
Medium *b*, High *a*	22	12	77	63	70	53
High *b*, Low *a*	53	34	59	42	48	32
DIF Effect Size (Area)						
.4	23	12	44	28	38	21
.6	36	23	65	49	59	41
.8	47	33	78	64	74	58
1.0	50	38	83	71	80	67
Mean Type I Error Rates						
Sample Size						
$N_R = 500$, $N_F = 200$	4.4	.8	7.8	2.2	7.5	1.8
$N_R = 500$, $N_F = 500$	4.7	1.1	8.7	2.6	8.4	2.3
$N_R = 1{,}000$, $N_F = 200$	4.5	1.0	8.2	2.5	8.5	2.2
$N_R = 1{,}000$, $N_F = 500$	5.6	1.4	9.1	2.7	8.9	2.3
Ability Distribution						
Equal	4.1	.9	7.0	1.8	6.1	1.4
Unequal	5.5	1.3	10.0	3.2	9.8	3.0
% DIF						
0%	4.9	1.0	8.1	2.2	7.5	2.0
10%	4.8	1.1	8.4	2.5	8.6	2.3
20%	4.8	1.0	8.6	2.6	8.1	2.2
Type of Item						
Low *b*, High *a*	5.2	1.0	10.5	2.5	9.9	2.4
Medium *b*, Low *a*	4.4	1.1	7.6	1.8	7.5	1.3
Medium *b*, High *a*	5.3	1.2	9.4	1.9	8.8	2.2
High *b*, Low *a*	4.3	1.3	6.5	1.5	6.4	1.9

items. The Type I error rates were slightly higher for CRO-SIB and LR, ranging up to approximately 8.6%. At $\alpha = .05$, the Type I error rates were within the nominal limits for MH (approximately 4.8%) and were higher than expected for CRO-SIB (8.1% to 8.6%) and LR (7.5% to 8.6%) for tests with 0%, 10%, or 20% DIF items.

Table 6
Mean Percent Detection Rates (Power) for the MH, CRO-SIB, and LR Procedures for Ability Distribution × N and for Ability Distribution × Type of Item for $\alpha = .05$ and $\alpha = .01$

	Equal Ability Distribution						Unequal Ability Distribution					
	MH		CRO-SIB		LR		MH		CRO-SIB		LR	
Factor	.05	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05	.01
Sample Size												
$N_R = 500$, $N_F = 200$	34	24	59	43	58	38	27	13	57	38	46	27
$N_R = 500$, $N_F = 500$	43	35	74	61	77	66	40	24	69	54	60	44
$N_R = 1{,}000$, $N_F = 200$	37	27	64	49	63	42	34	16	60	44	50	31
$N_R = 1{,}000$, $N_F = 500$	45	39	79	69	81	70	43	34	79	68	69	56
Type of Item												
Low *b*, High *a*	91	82	93	86	93	81	42	26	82	70	87	73
Medium *b*, Low *a*	6	3	44	27	48	30	23	10	50	32	39	22
Medium *b*, High *a*	3	1	81	68	84	68	41	23	73	58	57	37
High *b*, Low *a*	59	40	59	41	54	37	47	29	59	43	42	26

Effect of type of item. The results in Table 5 show that overall the detection rates for the four types of items were lowest for MH, whereas the similarity between the detection rates for CRO-SIB and LR was very high. The detection rates for CRO-SIB and LR were highest for low *b*, high *a* items (88% and 90% for $\alpha = .05$) followed by medium *b*, high *a* items (77% and 70% for $\alpha = .05$); for MH, these percentages were 66% and 22%. The detection rates for CRO-SIB and LR were lowest for medium *b*, low *a* items (47% and 44% for $\alpha = .05$) followed by high *b*, low *a* items (59% and 48% for $\alpha = .05$); corresponding results for MH were 15% and 53%.

The Type I error rates for MH were well within expected limits, and higher than expected for CRO-SIB and LR (10.5% and 9.9% for $\alpha = .05$). For all three procedures, the Type I error rates were higher for highly discriminating items.

Effect of DIF effect size. The detection rates for the three procedures steadily increased as area values increased from .4 to 1.0. The lowest detection rates were observed for MH, which ranged from 23% to 50% when the area value increased from .4 to 1.0 for $\alpha = .05$. For CRO-SIB, they ranged from 44% to 83%; for LR, they ranged from 38% to 80%.

Table 4 shows the results of 10 two-way interaction effects between the five factors. Therefore, care is needed in interpreting the main effects of the ANOVA in view of the significant two-way interactions between the factors. Tables 6 and 7 show mean detection rates for four of the significant interactions.

Effect of ability distribution × N. Table 6 shows that as N increased, the detection rates for all three procedures increased for the equal as well as the unequal ability distribution. The lowest detection rates were obtained for MH under both conditions. The detection rates for all three procedures were higher for the equal ability distribution than those for the unequal ability distribution. At $\alpha = .05$, as N increased the detection rates for the equal ability distribution increased from 34% to 45% for MH, from 59% to 79% for CRO-SIB, and from 58% to 81% for LR. For the unequal ability distribution, the detection rates increased from 27% to 43% for MH, from 57% to 79% for CRO-SIB, and from 46% to 69% for LR as N increased. The interaction between N and ability distribution showed a decrease in the detection rates for the unequal ability distribution of 7% for MH, 2% for CRO-SIB, and 12% for LR for $N_R = 500$ and $N_F = 200$, and 2% for MH, 0% for CRO-SIB, and 12% for LR for $N_R = 1{,}000$ and $N_F = 500$ from the detection rates for the equal ability distribution.

Effect of ability distribution × type of item. Table 6 also shows the interaction between ability distribution and type of item for the three procedures. Several trends were evident from the data in Table 6 at $\alpha = .05$.

1. The detection rates for all three procedures for the equal ability distribution were highest for items with low *b* and high *a*. At $\alpha = .05$, they were 91% for MH, and 93% for CRO-SIB and LR. Detection rates for the unequal ability distribution decreased by 49% for MH, 11% for CRO-SIB, and 6% for LR.
2. For items with high *b* and low *a* for the equal ability distribution, the detection rates were 59% for MH and CRO-SIB, and 54% for LR. The detection rates for the unequal ability distribution decreased by 12% for MH, remained the same at 59% for CRO-SIB, and decreased by 12% for LR.
3. For medium *b* items (for both low and high *a*s), the detection rates of MH for both distributions were very low compared to those of CRO-SIB and LR. For the equal ability distribution, the detection rates for medium *b*, low *a* items were 6% for MH, 44% for CRO-SIB, and 48% for LR. For the unequal ability distribution, the detection rates increased by 17% for MH and by 6% for CRO-SIB, and decreased by 9% for LR.
4. For the equal ability distribution, the detection rates for medium *b*, high *a* items were 3% for MH, 81% for CRO-SIB, and 84% for LR. For the unequal ability distribution, the detection rates increased by 38% for MH, decreased by 8% for CRO-SIB, and decreased by 27% for LR.

Thus, CRO-SIB and LR showed the highest detection rates for items with high *a*s and low *b*s. The detection rates for MH were highest for the unequal ability distribution for medium *b* items.

Regardless of the type of item, MH showed a power of only approximately 50% when the ability distribution was unequal. Therefore, MH appears to be not suitable for studying nonuniform DIF when the ability distributions are unequal. With the equal ability distribution, MH was effective in detecting nonuniform DIF only for items with high and low *b*s, a situation that corresponds to ordinal or unidirectional DIF.

Effect of percent of DIF × N. Table 7 shows the results of the interaction effects of $N \times$ by percent of items containing DIF. Again, Table 7 shows that as *N* increased, the detection rates for the three procedures increased whether tests had 10% or 20% DIF items. At $\alpha = .05$, as *N* increased the detection rates for MH increased from 31% to 49% for tests containing 10% and 20% DIF items. The detection rates for CRO-SIB increased from 58% to 78% when tests contained 10% DIF items and from 57% to 80% when tests contained 20% DIF items. For both MH and CRO-SIB, the interaction between *N* and % DIF was minimal for all *N*s (approximately 1% to 2%). As *N* increased, the detection rates for LR increased from 54% to 78% when tests contained 10% DIF items and from 49% to 73% when tests contained 20% DIF items. The interaction between *N* and % DIF for LR showed a decrease up to approximately 5% in the detection rates for tests containing 20% DIF items from tests containing 10% DIF items, depending on *N*.

Effect of percent of DIF × type of item. The results for the interaction effects of the % DIF × type of item (Table 7) show that the detection rates for CRO-SIB did not differ much for all types of items as the

Table 7
Mean Percent Detection Rates (Power) for the MH, CRO-SIB, and LR Procedures for % DIF × *N* and for % DIF × Type of Item for $\alpha = .05$ and $\alpha = .01$

	10% DIF Items						20% DIF Items					
	MH		CRO-SIB		LR		MH		CRO-SIB		LR	
Factor	.05	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05	.01
Sample Size												
$N_R = 500, N_F = 200$	31	18	58	41	54	34	31	19	57	40	49	31
$N_R = 500, N_F = 500$	41	29	71	57	71	58	41	30	72	58	67	52
$N_R = 1{,}000, N_F = 200$	36	22	63	47	58	38	35	22	61	47	55	35
$N_R = 1{,}000, N_F = 500$	49	37	78	68	78	67	49	37	80	69	73	60
Type of Item												
Low *b*, High *a*	64	51	89	76	91	78	69	57	88	79	89	76
Medium *b*, Low *a*	13	5	46	28	44	26	17	8	48	31	43	26
Medium *b*, High *a*	26	14	77	64	73	55	19	10	77	63	68	51
High *b*, Low *a*	54	35	61	43	52	35	52	34	57	40	45	28

percent of items containing DIF increased. For LR, the detection rates decreased for all types of items as the percent of items showing DIF decreased. For MH, the detection rates decreased for medium *b*, high *a* and high *b*, low *a* items and increased for low *b*, high *a* and medium *b*, low *a* items.

For example, at $\alpha = .05$, the detection rates for tests containing 10% DIF items were 64% for MH, 89% for CRO-SIB, and 91% for LR for low *b*, high *a* items. They increased by 5% for MH, decreased by 1% for CRO-SIB, and decreased by 2% for LR for tests containing 20% DIF items. For high *b*, low *a* items, the detection rates for tests containing 10% DIF items were 54% for MH, 61% for CRO-SIB, and 52% for LR. They decreased by 2% for MH, by 4% for CRO-SIB, and by 7% for LR for tests containing 20% DIF items. For medium *b*, low *a* items, the detection rates for tests containing 10% DIF items were 13% for MH, 46% for CRO-SIB, and 44% for LR. They increased by 4% for MH, by 2% for CRO-SIB, and by 1% for LR for tests containing 20% DIF items. For medium *b*, high *a* items, the detection rates for tests containing 10% DIF items were 26% for MH, 77% for CRO-SIB, and 73% for LR. They decreased by 7% for MH, remained the same for CRO-SIB, and decreased by 5% for LR for tests containing 20% DIF items.

The interaction between % DIF and type of item showed a marginal decrease of up to 4% for CRO-SIB for low *a* items in tests containing 20% DIF items from those containing 10% DIF items. For LR, there was an overall decrease ranging from 1% to 7% for tests with 20% DIF items from tests with 10% DIF items, depending on the type of item. MH showed an increase of approximately 5% and a decrease of approximately 2% to 7% in the detection rates, depending on the type of item.

These results indicate that although CRO-SIB and LR in general are able to identify a high percentage of nonuniform DIF items, their inflated Type I error rates call for an adjustment to the values at the desired significance levels. To investigate such an adjustment, the Type I error rates of CRO-SIB and LR were evaluated at nine significance levels—.05, .04, .03, .02, .01, .0075, .005, .0025, and .001—to determine the exact level of adjustment to the values at the desired level.

Table 8 presents the Type I error rates of CRO-SIB and LR statistics at nine significance levels. These results indicate that the Type I error rates varied across all three factors—*N*, ability distribution, and % DIF. Figure 1a demonstrates graphically the results presented in Table 8 for $N_R = 500$ and $N_F = 200$; Figure 1b displays the results for the two ability distributions, averaged across sample sizes. Figure 1a shows that the level of adjustment for $\alpha = .05$ is to set it to $\alpha = .03$ for both procedures. Figure 1b shows that the impact of the equal and unequal ability distributions on the Type I error rates of the two procedures was different. Although the Type I error rates were only slightly inflated for the two procedures for the equal ability distributions, they were much higher for the unequal ability distributions. For the equal ability distribution, the level of adjustment for $\alpha = .05$ is to set it to $\alpha = .034$ for CRO-SIB and to $\alpha = .04$ for LR (see Figure 1b). For the unequal ability distribution, the level of adjustment for $\alpha = .05$ is to set it to $\alpha = .022$ for CRO-SIB and to $\alpha = .025$ for the LR. These results appear to be a means to ensure that the Type I error rates are under control.

Discussion

The results indicated that overall there was high agreement between CRO-SIB and LR in detecting nonuniform DIF under most conditions. It is not surprising that MH was not capable of detecting nonuniform DIF under certain conditions because this procedure was designed to detect only uniform DIF. As expected, all three procedures were affected by sample size. The detection rates for all three procedures increased when sample size increased. Given that the reference group had a minimum sample size of 500, the results of this study indicate that detection rates are affected by small focal group sample sizes. When the focal group sample size increased from 200 to 500, the detection rates increased. The power of CRO-SIB and LR in detecting nonuniform DIF was as high as 75%, on average, for a focal group sample size of approximately 500. Because this study investigated only four combinations of sample size, more research is needed in this area. The ratio of the reference to the focal group sample size should be taken into consideration.

Table 8
Type I Error Rates of CRO-SIB and LR Computed at Nine Significance Levels

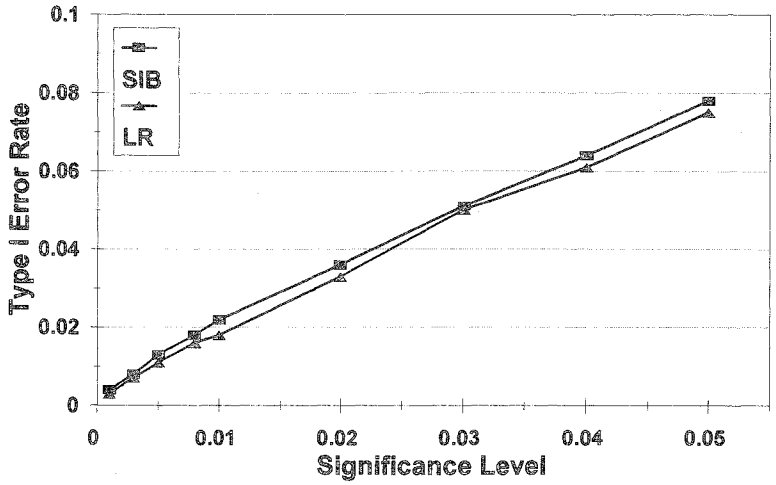
Statistic and Factor	Significance Level (α)								
	.05	.04	.03	.02	.01	.0075	.005	.0025	.001
CRO-SIB									
Sample Size									
$N_R = 500, N_F = 200$	7.8	6.4	5.1	3.6	2.2	1.8	1.3	.8	.4
$N_R = 500, N_F = 500$	8.7	7.1	5.8	4.3	2.6	2.0	1.3	.9	.5
$N_R = 1{,}000, N_F = 200$	8.2	6.6	5.3	3.8	2.5	1.9	1.5	1.0	.5
$N_R = 1{,}000, N_F = 500$	9.1	7.2	5.9	4.1	2.7	2.0	1.5	1.0	.5
Ability Distribution									
Equal	7.0	5.8	4.5	3.1	1.8	1.4	1.0	.6	.3
Unequal	10.0	7.8	6.4	4.8	3.2	2.5	1.9	1.2	.6
Percent of DIF									
0%	8.1	6.5	5.1	3.7	2.2	1.6	1.1	.6	.3
10%	8.4	6.9	5.6	4.0	2.5	2.0	1.4	.9	.5
20%	8.5	6.7	5.3	3.8	2.6	1.9	1.4	.8	.4
LR									
Sample Size									
$N_R = 500, N_F = 200$	7.5	6.1	5.0	3.3	1.8	1.6	1.1	.7	.3
$N_R = 500, N_F = 500$	8.4	6.9	5.2	4.1	2.2	1.7	1.0	.8	.4
$N_R = 1{,}000, N_F = 200$	8.5	7.0	5.3	4.0	2.4	1.7	1.1	.9	.4
$N_R = 1{,}000, N_F = 500$	8.9	6.9	5.1	3.8	2.3	1.8	1.2	.9	.4
Ability Distribution									
Equal	6.1	5.1	4.0	2.8	1.4	1.1	.8	.4	.2
Unequal	9.8	7.4	6.1	4.2	3.0	2.3	1.6	1.0	.4
Percent of DIF									
0%	7.5	5.8	4.8	3.3	2.0	1.3	.9	.4	.2
10%	8.6	6.0	5.1	3.6	2.3	1.5	1.0	.6	.3
20%	8.1	5.7	5.0	3.5	2.2	1.5	1.0	.5	.3

The results also suggest that DIF effect size can have a significant effect on DIF detection procedures irrespective of the size of reference and focal groups and the ratio of these sample sizes. For all three procedures, the detection rates steadily increased when DIF effect sizes specified in terms of the areas between the IRFs for the two groups increased from .4 to 1.0. The lowest detection rates occurred for MH—varying between 23% and 50%. Practitioners should be aware that items showing very small amounts of DIF may go undetected, especially when the sample is small. However, it can be argued that in such cases the DIF may be so small that it would make little practical difference.

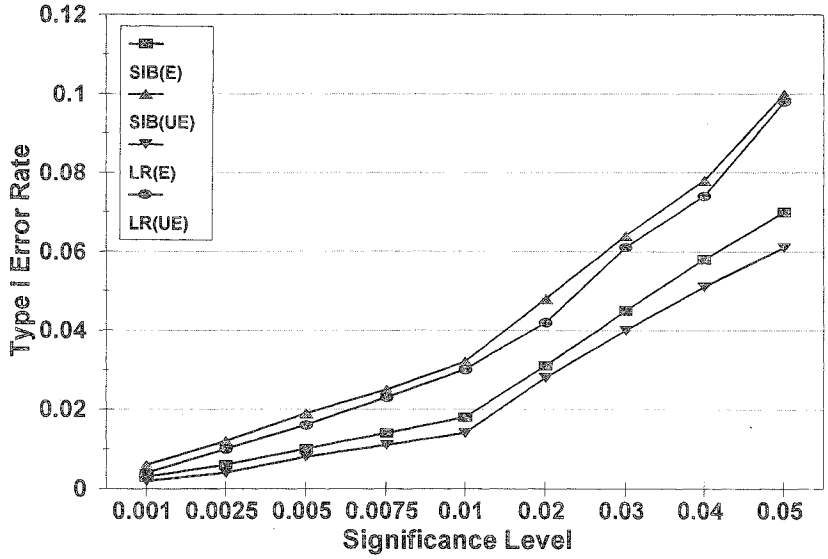
The results supported the findings of Rogers & Swaminathan (1993a); that is, the difficulties and discriminations of the items that comprise the test significantly influenced the detection rates of DIF detection procedures. Their study comparing MH and LR showed that MH was not capable of detecting nonuniform DIF when the interaction was disordinal or nondirectional (i.e., when the IRFs of the two groups crossed in the middle of the ability distribution). Disordinal or nondirectional interactions occur with items of average difficulty. The MH statistic is a signed statistic and, thus, is sensitive to the direction of DIF. When the direction of DIF changes in the middle of the ability score distribution, negative differences in one part of the score distribution cancel against the positive differences in the other part. Therefore, nonuniform DIF items of this form will not be detected by MH. CRO-SIB was as powerful as the LR procedure in detecting ordinal and disordinal interactions under most of the studied conditions. For the two types of items included in this study for which the interactions were ordinal (when the IRFs for the two groups crossed at the lower or upper end of the ability distribution), the performance of MH was comparable with the other two procedures.

Figure 1
Type I Error Rates for the SIB and LR Procedures

a. $N_R = 500$ and $N_F = 200$



b. Equal (E) and Unequal (UE) Ability Distributions



In general, the detection rates for CRO-SIB and LR were highest for high *a* items with low *b* followed by medium *b* items. Low *a* items with medium *b* were least detected. For MH, the most significant factor to determine its capability to detect nonuniform DIF appeared to be the type of item. Although its performance appeared to be comparable with the other two procedures in detecting DIF in low and high *b* items, MH has limited use in the detection of DIF in average *b* items. However, it appears that DIF in such items can be adequately detected by CRO-SIB and LR.

The proportion of items showing DIF did not have a large impact on the DIF detection rates of MH and CRO-SIB; however, it did minimally affect the detection rates of LR (approximately 4%). This may be due to the two-stage procedure adopted in computing the MH statistic. Items identified as DIF in the first computations were removed when forming the score groups for computing the DIF statistics for the second time. The two-stage procedure was not used for CRO-SIB and LR, and it is likely that the results would have improved for both procedures if this had been used.

The Type I error rates were within nominal limits for MH. They were higher than expected for CRO-SIB and LR, with CRO-SIB results showing an overall increase of approximately 1% over LR results. In general, there appeared to be an increase in Type I error rates for the three procedures when the ability distribution differences increased or proportion of items containing DIF increased.

The results also showed that CRO-SIB and LR were equally effective in detecting nonuniform DIF in test items. However, the Type I error rates for both procedures were higher than the nominal level and, therefore, require an adjustment. This study indicated that the levels of adjustment varied with different conditions. The exact level of adjustment can be determined by evaluating the Type I error rates at a number of significance levels. The desired significance level then can be set to the adjusted significance level for the condition investigated. MH appeared to have limited use in the detection of nonuniform DIF items that crossed in the middle of the ability range for the equal ability distribution. For the unequal ability distribution, its power was limited for all types of items.

In general, the results indicate that with an adjustment in the α level, either the CRO-SIB procedure or the LR procedure can be used routinely for DIF detection. CRO-SIB is noniterative and simple to implement. However, LR is a general procedure and can be implemented readily using computer packages such as SPSS (SPSS, 1993) and SAS/STAT (SAS, 1993).

References

Ackerman, T. A. (1992, April). *An investigation of the relationship between reliability, power, and Type I error rates of the Mantel-Haenszel and the simultaneous item bias procedures.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Berk, R. A. (Ed.). (1982). *Handbook of methods for detecting test bias.* Baltimore MD: Johns Hopkins University Press.

Bock, R. D. (1975). *Multivariate statistical methods in behavioral research.* New York: McGraw-Hill.

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

Hambleton, R. K., & Rogers, H. J. (1989). Detecting potentially biased test items; comparison of IRT area and Mantel-Haenszel methods. *Applied Measurement in Education, 2,* 313–334.

Hambleton, R. K., & Rovinelli, R. (1973). A FORTRAN IV program for generating examinee response data from logistic test models. *Behavioral Science, 18,* 73–74.

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

Kingston, N., Leary, L., & Wightman, L. (1988). *An exploratory study of the applicability of item response theory methods to the Graduate Management Admission Test* (GMAT Occasional Papers). Princeton NJ: Graduate Management Admission Council.

Li, H., & Stout, W. F. (1993, April). *A new procedure for detection of crossing DIF/bias.* Paper presented at the annual meeting of the American Educational Research Association, Atlanta.

Li, H., & Stout, W. F. (1994a). *CSIBTEST: A FORTRAN-V program for computing the simultaneous item bias DIF statistics* [Computer program]. Urbana-Champaign: University of Illinois, Department of Statistics.

Li, H., & Stout, W. F. (1994b, April). *Detecting crossing item bias/DIF: Comparison of logistic regression and crossing SIBTEST procedures.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans LA.

Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. (1981). An investigation of item bias in a test of reading comprehension. *Applied Psychological Measurement, 5,* 159–173.

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1992). The effect of sample size on the functioning of the

Mantel-Haenszel statistic. *Educational and Psychological Measurement, 52,* 443–451.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics, 7,* 105–118.

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

Narayanan, P., & Swaminathan, H. (1994). Performance of the Mantel-Haenszel and simultaneous item bias procedures for detecting differential item functioning. *Applied Psychological Measurement, 18,* 315–338.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 53,* 495–502.

Rogers, H. J., & Swaminathan, H. (1993a). A comparison of the logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17,* 105–116.

Rogers, H. J., & Swaminathan, H. (1993b). *DICHOdif: A FORTRAN-V program for computing the Mantel-Haenszel and the logistic regression DIF statistics* [Computer program]. New York: Columbia University, Teacher's College.

Roussos, L. A., & Stout, W. F. (1993, April). *Simulation studies of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel Type 1 error performance.* Paper presented at the annual meeting of the American Educational Research Association, Atlanta.

SAS Institute, Inc. (1993). *SAS/STAT user's guide* (Version 6.0) [Computer software]. Cary NC: Author.

Shealy, R., & Stout, W. F. (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,* 159–174.

SPSS, Inc. (1993). *SPSS for windows* (Version 6.0) [Computer software]. Chicago: Author.

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

Author's Address

Send requests for reprints or further information to Pankaja Narayanan, Mail Stop 53-L, Educational Testing Service, Rosedale Road, Princeton NJ 08541, U.S.A. Email: pnarayanan@ets.org.