Performance of the Mantel-Haenszel and Simultaneous Item Bias Procedures for Detecting Differential Item Functioning

Pankaja Narayanan and H. Swaminathan
University of Massachusetts at Amherst

Two nonparametric procedures for detecting differential item functioning (DIF)—the Mantel-Haenszel (MH) procedure and the simultaneous item bias (SIB) procedure—were compared with respect to their Type I error rates and power. Data were simulated to reflect conditions varying in sample size, ability distribution differences between the focal and reference groups, proportion of DIF items in the test, DIF effect sizes, and type of item. 1,296 conditions were studied. The SIB and MH procedures were equally powerful in detecting uniform DIF for equal ability distributions. The SIB procedure was more powerful than the MH procedure in detecting DIF for unequal ability distributions. Both procedures had sufficient power to detect DIF for a sample size of 300 in each group. Ability distribution did not have a significant effect on the SIB procedure but did affect the MH procedure. This is important because ability distribution differences between two groups often are found in practice. The Type I error rates for the MH statistic were well within the nominal limits, whereas they were slightly higher than expected for the SIB statistic. Comparisons between the detection rates of the two procedures were made with respect to the various factors.

Index terms: differential item functioning, Mantel-Haenszel statistic, power, simultaneous item bias statistic, SIBTEST, Type I error rates.

In recent years, the concern over the issue of differential item functioning (DIF) in standardized achievement and ability tests has resulted in the development of a variety of statistical methods for detecting DIF. Item response theory (IRT) provides a general framework for studying DIF. However, IRT-based procedures require large sample sizes, a condition that is often difficult to meet in practice. Because of this problem, non-IRT procedures are frequently used to detect DIF.

Some of the currently available techniques for detecting DIF are the Mantel-Haenszel (MH) procedure (Holland & Thayer, 1988), the standardization procedure (Dorans & Kulick, 1986), the simultaneous item bias procedure [SIBTEST (henceforth referred to as SIB); Shealy & Stout, 1991], and the logistic regression procedure (Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990). The MH procedure has been shown to be one of the most effective methods for detecting DIF (Hambleton & Rogers, 1989; Mazor, Clauser, & Hambleton, 1992; Raju, Bode, & Larsen, 1989; Shealy & Stout, 1993).

MH and SIB share a common framework. Both procedures are nonparametric, and therefore do not require model calibration (Ackerman & Evans, 1992). They also provide tests of significance, are computationally simple, and inexpensive. The MH and SIB procedures typically use the number-correct score as the conditioning variable to form groups of examinees of comparable ability. For two groups matched on K score categories, the MH procedure compares the odds of success for the reference and focal groups. 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. Instead of matching on total score, SIB allows the user to select the matching subtest, called the "valid subtest." For examinees who are matched on K valid subtest score categories, SIB compares the average proportion correct on the "suspect" subtest for the reference and

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 18, No. 4, December 1994, pp. 315–328
© Copyright 1994 Applied Psychological Measurement Inc.
0146-6216/94/040315-14$1.95

the focal group examinees. In addition, the SIB procedure, unlike the MH procedure, can simultaneously evaluate DIF present in several test items.

Zwick (1990) argued that the MH procedure may have a higher Type I error rate than expected when the probability of a correct response to an item can be described by a two- or a three-parameter item response model rather than a one-parameter model. Using simulated data, Shealy & Stout (1993) and Roussos & Stout (1993) showed that, although in general the MH and the SIB procedures yielded comparable Type I error rates, under certain extreme conditions the MH procedure yielded higher Type I error rates than the SIB procedure. Ackerman & Evans (1992) demonstrated that in the case in which multiple items exhibit DIF, the SIB procedure, with its emphasis on the selection of a valid subtest for matching the examinees, performed better than the MH procedure when the total score was used as the matching criterion. However, this may be due to the choice of the matching criterion rather than the choice of the procedure.

Although considerable research has been done on the MH procedure, relatively little research has been conducted on the SIB procedure. Therefore, this study compared the Type I error rates and the power of the MH and SIB procedures under a variety of conditions to identify the conditions under which each procedure is optimal for detecting DIF.

Description of the DIF Statistics

The MH 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. 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.

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

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

The null DIF hypothesis is that the odds of getting the item correct at a given score level j are the same for the reference and the focal group at all K levels of the matching criterion. The null and alternative constant odds ratio hypothesis at score level j can be expressed as

$$\mathrm{H}_0\text{: } \left[\pi_{\mathrm{R}j}/\left(1-\pi_{\mathrm{R}j}\right)\right]=\left[\pi_{\mathrm{F}j}/\left(1-\pi_{\mathrm{F}j}\right)\right] \qquad j=1,2,\ldots,k\,, \tag{1}$$

and

$$\mathrm{H}_\mathrm{A}\text{: } \left[\pi_{\mathrm{R}j}/\left(1-\pi_{\mathrm{R}j}\right)\right]=\alpha\left[\pi_{\mathrm{F}j}/\left(1-\pi_{\mathrm{F}j}\right)\right] \qquad j=1,\ 2,\ \ldots,\ k,\quad \alpha\neq 1\,, \tag{2}$$

where $\pi_{\mathrm{R}j}$ is the probability that a reference group (R) examinee with total score j will answer the studied item correctly, and $\pi_{\mathrm{F}j}$ 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

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

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

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

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

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

and

$$\text{Var}\left(A_j\right) = \frac{N_{\text{R}j} N_{\text{F}j} N_{1.j} N_{0.j}}{\left(N_{..j}\right)^2 \left(N_{..j} - 1\right)}. \tag{6}$$

The SIB Procedure

The SIB procedure (Shealy & Stout, 1991) 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 test 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 score 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. The hypothesis of interest is H_0: $\beta_{\text{U}} = 0$ versus H_{A}: $\beta_{\text{U}} > 0$. H_{A} is a one-sided test to specifically test for DIF against the focal group.

Let

$$X = \sum_{i=1}^{n} U_i \tag{7}$$

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

$$Y = \sum_{i=n+1}^{N} U_i \tag{8}$$

be the total score on the studied subtest. Let $\overline{Y}_{\text{R}k}$ and $\overline{Y}_{\text{F}k}$ be the average score on the suspect subtest for all examinees in the reference and the focal groups, respectively, attaining a valid subtest score $X = k$ ($k = 0, 1, 2, ..., n$). Because $\left(\overline{Y}_{\text{R}k} - \overline{Y}_{\text{F}k}\right)$ 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 distributions of the reference and the focal groups, even in the case of no DIF $\left(\overline{Y}_{\text{R}k} - \overline{Y}_{\text{F}k}\right)$ 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 $\overline{Y}_{\text{R}k}$ and $\overline{Y}_{\text{F}k}$. [For more details on the classical test theory and IRT-based justification for the regression correction, refer to Shealy & Stout (1993).] It follows that an estimate $\hat{\beta}_{\text{U}}$ of β_{U} is

$$\hat{\beta}_{\text{U}} = \sum_{k=0}^{n} \hat{p}_k \left(\overline{Y}_{\text{R}k} - \overline{Y}_{\text{F}k}\right), \tag{9}$$

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 DIF is given by

$$B_{\text{U}} = \hat{\beta}_{\text{U}} / \hat{\sigma}\left(\hat{\beta}_{\text{U}}\right), \tag{10}$$

where $\hat{\sigma}(\hat{\beta}_U)$ is the estimated standard error of β_U. The expression for $\hat{\sigma}(\hat{\beta}_U)$ (Shealy & Stout, 1993, p. 169) is

$$\hat{\sigma}(\hat{\beta}_U) = \left\{\sum_{k=0}^{n} \hat{p}_k^2 \left[\frac{1}{J_{Rk}}\hat{\sigma}^2(Y|R,k) + \frac{1}{J_{Fk}}\hat{\sigma}^2(Y|F,k)\right]\right\}^{1/2}, \quad (11)$$

where J_{Rk} and J_{Fk} are the number of examinees in the reference and focal groups with the same valid score $x = k$.

The test statistic B_U has an approximate N(0,1) distribution when no DIF is present. The null hypothesis of no DIF is rejected if the value of B_U exceeds the upper $100(1 - \alpha)$th percentile point of the standard normal distribution.

Method

Examinee response data were simulated under a variety of conditions, with each dataset accommodating prespecified levels of a number of different factors that might have an effect on DIF detection rates or the power of the MH and SIB procedures. This study was confined to the investigation of uniform DIF because the MH procedure is designed to detect only uniform DIF. Because the number of items was not a factor manipulated in this study, a test length of 40 items was used to investigate the capability of MH and SIB to detect DIF in a "short" test.

Manipulated Variables

Five factors were manipulated: sample size, ability distribution differences, proportion of items containing DIF, DIF effect size, and type of item.

Sample size. One factor of interest was sample size in the focal and reference groups. Research conducted on the power of the MH procedure has indicated that DIF detection rates increase with increased sample size (Mazor et al., 1992; Rogers, 1989; Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990). The three reference group sample sizes (N_R = 300, N_R = 500, and N_R = 1,000) were crossed with the three focal group sample sizes (N_F = 100, N_F = 200, and N_F = 300) to produce nine sample size combinations.

Ability distribution differences. A second factor of interest was the ability distribution differences between the two groups. Mazor et al. (1992) studied the effects on the MH procedure when two groups were sampled from equal and unequal distributions. They recommended that when comparing groups of differing abilities large sample sizes should be used. Shealy & Stout (1993) showed that both MH and SIB displayed good adherence to the nominal level of significance even for differences in ability as large as 1 standard deviation (SD).

The impact of the differences in underlying ability distributions was investigated by examining three different conditions that were studied by Shealy & Stout (1993). In the first case, the mean of the ability distributions for the two groups was set equal to 0.0, and the SD was set equal to 1.0. This will be referred to as the equal ability distribution.

In the second condition, the mean was set equal to 0.0 and .5 for the reference and focal groups, respectively, with both SDs set equal to 1.0. This will be referred to as the $U_{.5}$ ability distribution. Ability distributions that differed by .5 SDs simulated the case in which there is not a very substantial between-group difference.

In the third condition, the mean was set equal to 0.0 and -1.0 for the reference and the focal groups, respectively, again with both SDs set equal to 1.0. This will be referred to as the $U_{1.0}$ ability distribution. Ability distributions that differed by 1 SD simulated the case in which there is a substantial between-group difference.

Proportion of DIF items. A third factor of interest was the proportion of items exhibiting DIF. In general, a longer test is likely to produce more reliable scores, resulting in more reliable ability estimates. On the other hand, increasing the proportion of items exhibiting DIF will produce ability estimates that will be less reliable. When the ability estimates are less reliable, matching will be less accurate. Therefore, the power of the DIF procedures is likely to decrease. Ackerman & Evans (1992) investigated the issue of reliability and its impact on the MH and SIB statistics. Their results suggested that the power of both statistics increased moderately as reliability increased and substantially as sample sizes increased. However, power decreased for a combination of small sample size and low or high reliability.

To study the effect of the proportion of items exhibiting DIF, tests were simulated with either 10% or 20% of the items showing DIF. In practice it is not unusual for 10% to 15% of the items in a standardized achievement test to exhibit DIF (Clauser, 1993). The 20% proportion of DIF items was included to represent the "worst case" scenario.

DIF effect size. DIF effect size or the amount of DIF contained in an item is the fourth factor that is likely to have an effect on the DIF detection procedures. As DIF effect size increases, the detection rates of the two procedures are expected to increase as well.

DIF effect sizes were determined within an IRT framework. Within this framework, DIF exists if the item response functions (IRFs) for the two groups are not the same. Therefore, the difference between the IRFs for the two groups can be used as a measure of DIF effect size. If the difference between the IRFs is large, then DIF effect size is expected to be large; if the difference between the IRFs is small, then DIF effect size is expected to be small. Swaminathan & Rogers (1990) used the area between the focal and reference group IRFs as an operational measure of DIF effect size. In their study, they investigated area values ranging from .2 to .8.

The area between the IRFs for the two groups was used here to quantify the size of DIF. The areas between the IRFs were computed using equations given by Raju (1988). Four levels of DIF effect size were selected, equal to the area values .4, .6, .8, and 1.0. Area values in this range reflected DIF effect sizes ranging from a small amount of DIF (.4) to a fairly large amount of DIF (1.0).

Type of item. Uniform DIF was simulated by keeping the discrimination parameters (a) for the two groups the same, but varying the difficulty parameters (b) for the two groups. 24 items showing uniform DIF were obtained by varying the level of the common a parameter (low, medium, high), the level of the bs for the two groups (low, medium, high), and DIF effect size (area values of .4, .6, .8, and 1.0). Six types of item were studied: (1) low b, medium a; (2) low b, high a; (3) medium b, low a; (4) medium b, high a; (5) high b, low a; and (6) high b, medium a. (Other combinations of a and b did not yield area differences that were meaningful; therefore, these combinations were not studied.) The pseudo-guessing parameters (c) for the 24 DIF items were set equal to .20. The item parameters for the DIF items are shown in Table 2.

To simulate a 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 24 DIF items into six 40-item tests. Similarly, in order to simulate 20% of the items showing DIF (i.e., eight items), the 24 DIF items were distributed into three 40-item tests. The nonDIF items were the same in all the tests and did not vary across conditions. Item parameter values for the nonDIF items were randomly selected from published item parameter values from an administration of the Graduate Management Admissions Test (Kingston, Leary, & Wightman, 1988). The item parameter values for the nonDIF items are shown in Table 3.

Summary. DIF was analyzed for datasets simulated for nine combinations of sample size, three levels of ability distribution differences, two levels of proportion of DIF items, four levels of DIF effect size, and six item types. 1,296 conditions were studied ($9 \times 3 \times 2 \times 4 \times 6$). The data were replicated 100 times for each condition.

Data Generation and Analysis

Data were generated according to the three-parameter logistic model using the program DATAGEN (Hambleton & Rovinelli, 1973) to investigate the capability of the SIB and MH procedures to identify the 24 uniform DIF items described above. The χ^2_{MH} DIF statistic values for the MH procedure were obtained using the program MHBIAS (Rogers, 1991). The SIB DIF statistic values were obtained using the program SIBTEST (Shealy, Stout, & Roussos, 1991).

In computing the MH and SIB DIF statistics, a two-stage procedure recommended by Holland & Thayer (1988) was adopted. In the first stage, the total score

Table 2
Item Parameters Used to Generate Items with DIF

Item Type and Item Number	DIF Effect Size	b_R	b_F	a_R	a_F
Low *b*, Medium *a*					
1	.4	−1.80	−1.28	.90	.90
2	.6	−1.92	−1.14	.90	.90
3	.8	−2.04	−1.01	.90	.90
4	1.0	−2.16	−.88	.90	.90
Low *b*, High *a*					
5	.4	−1.80	−1.28	1.25	1.25
6	.6	−1.92	−1.14	1.25	1.25
7	.8	−2.04	−1.01	1.25	1.25
8	1.0	−2.16	−.88	1.25	1.25
Medium *b*, Low *a*					
9	.4	−.26	.26	.50	.50
10	.6	−.39	.39	.50	.50
11	.8	−.51	.51	.50	.50
12	1.0	−.64	.64	.50	.50
Medium *b*, High *a*					
13	.4	−.26	.26	1.25	1.25
14	.6	−.39	.39	1.25	1.25
15	.8	−.51	.51	1.25	1.25
16	1.0	−.64	.64	1.25	1.25
High *b*, Low *a*					
17	.4	1.28	1.80	.50	.50
18	.6	1.14	1.92	.50	.50
19	.8	1.01	2.04	.50	.50
20	1.0	.88	2.16	.50	.50
High *b*, Medium *a*					
21	.4	1.28	1.80	.90	.90
22	.6	1.14	1.92	.90	.90
23	.8	1.01	1.24	.90	.90
24	1.0	.88	2.16	.90	.90

based on all the items was used as the matching criterion to group the examinees, and items showing DIF were identified using the MH and the SIB procedures. In the second stage, items showing DIF (with the exception of the studied item for the MH procedure) were excluded from the calculation of total score used to group examinees. Then the MH and SIB analyses were repeated.

The power (percent of DIF items correctly identified as DIF) and Type I error rates (percent of nonDIF items falsely identified as DIF) of the MH and SIB statistics were evaluated at the $\alpha = .05$ and $\alpha = .01$ levels of statistical significance. An ANOVA was performed to determine the effects of the five conditions on the performance of the MH and SIB statistics. The dependent variable was the number of times the items were identified as DIF in 100 replications of the data. The independent variables were the five different conditions that were manipulated in the study.

Results

Table 4 shows the ANOVA results for the detection rates across all conditions for the SIB and MH statistics. The DIF detection rates of the MH and SIB procedures are shown in Tables 5 and 6.

Table 4 shows that for both SIB and MH, *N*, the percent of items containing DIF (% DIF), DIF effect size, and the type of item had significant main effects at $\alpha = .05$. For the SIB procedure, differences in the means of the ability distribution did not have a significant main effect, but they did for the MH

Table 3
Item Parameters for the Non-DIF Items

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

procedure.

The same two-way interaction effects were significant for both procedures: *N* × ability distribution differences, *N* × type of item, *N* × DIF effect size, ability distribution differences × type of item, ability distribution differences × DIF effect size, and type of item × DIF effect size. For both procedures, there was no interaction effect between % DIF and other factors. (All higher order interaction terms were grouped as error terms.)

Ability Distribution Differences

Table 5 shows the mean percent of items correctly identified as differentially functioning for the equal ability distribution for all conditions. Table 6 shows the mean percent of items correctly identified as differentially functioning for the $U_{.5}$ and $U_{1.0}$ ability distributions for all conditions. The main findings are summarized below.

Effect of sample size. For the equal, $U_{.5}$, and $U_{1.0}$ ability distributions, the mean percent detection rates for the two procedures showed a steady increase as the sample size increased (see Tables 5 and 6). In most cases, the SIB procedure identified a slightly higher percentage of DIF items than the MH procedure for the $U_{.5}$ and $U_{1.0}$ ability distributions (see Table 6).

Proportion of DIF items. There was an overall decrease of approximately 1% to 5% for the two procedures as the proportion of items showing DIF increased from 10% to 20%. For the equal distribution (Table 5), for $N_R = 500$, $N_F = 100$, the decrease for both procedures was 1%; for $N_R = 300$, $N_F = 200$, the decrease was 4% for SIB and 5% for MH. For the $U_{.5}$ distribution (Table 6), for $N_R = 1{,}000$, $N_F = 100$, the decrease was 5% for SIB and 4% for MH. For the $U_{1.0}$ distribution, for $N_R = 1{,}000$, $N_F = 100$, the decrease for both procedures was 2%. In general, the detection rates for both procedures showed a similar pattern for 10% and 20% DIF.

DIF effect sizes. For the equal as well as the $U_{.5}$ and $U_{1.0}$ ability distributions (Tables 5 and 6), the mean percent detection rates for the two procedures steadily increased for increased DIF effect sizes for all sample sizes.

Type of item. Several trends are evident from the data in Tables 4–6.

1. For the equal ability distribution (Table 5), the detection rates for the two procedures were highest for medium *b*, high *a* items followed by low *b*, high *a* items. The lowest detection rates were obtained for high *b*, low *a* items followed by high *b*, medium *a* items. In general, as *a* increased, the power of the two DIF procedures increased.
2. The results for the $U_{.5}$ and $U_{1.0}$ ability distributions (Table 6) reveal that for medium *b*, high *a*

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

	SIB		MH	
Factor	*F*	*p*	*F*	*p*
Main Effects				
N	273.60	<.001*	209.95	<.001*
Ability Distribution	.65	.520	260.50	<.001*
% DIF	31.95	<.001*	39.49	<.001*
Type of Item	1,737.39	<.001*	2,878.89	<.001*
DIF Effect Size	1,958.71	<.001*	1,857.50	<.001*
Interaction Effects				
N × Ability Distribution	3.32	<.001*	2.83	<.001*
N × % DIF	.30	.992	.22	.986
N × Type of Item	10.27	<.001*	6.84	<.001*
N × DIF Effect Size	5.63	<.001*	3.03	<.001*
Ability Distribution × % DIF	.03	.975	.02	.980
Ability Distribution × Type of Item	76.64	<.001*	184.12	<.001*
Ability Distribution × DIF Effect Size	42.58	<.001*	38.52	<.001*
% DIF × Type of Item	.99	.423	1.73	.124
% DIF × DIF Effect Size	1.93	.123	.69	.560
Type of Item × DIF Effect Size	69.62	<.001*	73.73	<.001*

*Significant at $\alpha = .05$.

items, the detection rates for the two procedures were comparable with those obtained with equal ability distributions. For low *b* items, the detection rates for both procedures were better than those obtained with the equal ability distribution regardless of the level of *a*. The detection rates for high *b* items were lower for both procedures than those obtained with the equal ability distribution regardless of the level of *a*.

3. A comparison of the detection rates of the two procedures showed that for medium *b*, low *a* items and the $U_{.5}$ distribution, the detection rates were 72% for SIB and 69% for MH (a difference of 4%). For the $U_{1.0}$ distribution, the detection rates were 77% for SIB and 64% for MH (a difference of 13%).
4. For high *b*, low *a* items, the detection rates were 58% for SIB and 56% for MH for the equal ability distribution. For the U_5 distribution, the detection rates decreased by approximately 7% for SIB and 15% for MH from the detection rates for the equal ability distribution. The detection rates for the $U_{1.0}$ distribution decreased by approximately 8% for SIB and 30% for MH from the detection rates for the equal ability distribution.
5. For high *b*, medium *a* items, the detection rates were 66% for SIB and 64% for MH for the equal ability distribution. For the U_5 distribution, the detection rates decreased by approximately 11% for SIB and 22% for MH from the detection rates for the equal ability distribution. The detection rates for the $U_{1.0}$ distribution decreased by approximately 22% for SIB and 45% for MH from the detection rates for the equal ability distribution.
6. Overall, the SIB procedure was able to identify a higher percentage of items for U_5 and $U_{1.0}$ than the MH procedure. For certain item types, SIB was able to detect approximately 10% to 25% more items than MH when the ability distributions were unequal. For example, for high *b*, low *a* items and the U_5 distribution, the detection rate for SIB was 51%, whereas it was 41% for MH (a decrease of about 10%). For high *b*, medium *a* items, the detection rate for SIB was 55% versus 42% for MH (a decrease of about 13%). For the $U_{1.0}$ distribution, for high *b*, low *a* items, the detection rate for SIB was 50% and 26% for MH (a decrease of about 24%). For high *b*, medium *a* items, the detection rate for SIB was 44% and 19% for MH (a decrease of about 25%).

Table 5
Mean Percent Detection Rates (Power) for the SIB and MH Procedures for the Equal Ability Distribution Under all Conditions for α = .05 and α = .01

	10% DIF				20% DIF			
	SIB		MH		SIB		MH	
Factor	.05	.01	.05	.01	.05	.01	.05	.01
Sample Size								
$N_R = 300$, $N_F = 100$	62	45	62	47	60	43	60	44
$N_R = 300$, $N_F = 200$	78	67	77	64	74	61	72	58
$N_R = 300$, $N_F = 300$	84	72	82	70	81	70	78	66
$N_R = 500$, $N_F = 100$	62	46	64	50	61	48	63	48
$N_R = 500$, $N_F = 200$	81	69	80	69	79	68	77	65
$N_R = 500$, $N_F = 300$	87	79	87	76	83	72	84	76
$N_R = 1{,}000$, $N_F = 100$	66	49	69	55	65	48	67	52
$N_R = 1{,}000$, $N_F = 200$	84	73	85	74	82	70	82	71
$N_R = 1{,}000$, $N_F = 300$	90	82	90	82	88	78	88	79
Type of Item								
Low *b*, Medium *a*	85	71	85	56	81	68	83	72
Low *b*, High *a*	88	76	89	81	85	75	86	79
Medium *b*, Low *a*	73	59	73	59	70	55	70	54
Medium *b*, High *a*	95	90	95	93	93	87	94	88
High *b*, Low *a*	58	40	56	36	55	37	51	34
High *b*, Medium *a*	66	52	64	48	66	50	63	45
DIF Effect Size (Area)								
.4	50	32	49	32	46	27	45	28
.6	75	59	76	61	72	56	72	56
.8	88	79	88	78	87	77	87	76
1.0	95	89	95	90	95	88	94	87

False Positive Rates

False positive or Type I error rates (number of nonDIF items falsely identified as DIF) for the two procedures were determined; Table 7 shows the means for the equal ability distribution and Table 8 shows means for the U_5 and $U_{1.0}$ ability distributions.

Sample size. Sample size did not affect Type I error rates for either procedure for the equal, U_5, and $U_{1.0}$ distributions. Overall, the SIB procedure had slightly higher Type I error rates than the MH procedure. For $N_R = 300$, $N_F = 100$ and the equal ability distribution, the Type I error rates were .062 for SIB and .037 for MH. For the U_5 distribution, they were .061 for SIB and .037 for MH; and for the $U_{1.0}$ distribution, they were .068 for SIB and .042 for MH. Roussos & Stout (1993), however, found that in one instance ($a = 2.5, b = -1.5$), the MH procedure yielded higher Type I error rates than the SIB procedure for $U_{1.0}$. In this condition, the Type I error rate of MH was .101 and .021 for SIB. The Type I error rates for the MH procedure for the conditions simulated in this study are similar to those found by others (Clauser, 1993; Rogers & Swaminathan, 1993). One possible explanation for Roussos's findings is that the *a* parameter used in his study was unrealistically high.

For the equal and U_5 ability distributions (Tables 7 and 8), at the $\alpha = .05$ and $\alpha = .01$ levels of significance, the Type I error rates for the MH procedure were the same as the nominal level for all sample sizes. Overall, the Type I error rates obtained for the SIB procedure were slightly higher than the nominal level. For the $U_{1.0}$ ability distribution (Table 8), the Type I error rates were inflated for both procedures; the inflation was slightly higher for the SIB procedure than for the MH procedure.

Proportion of DIF items. The Type I error rates for the MH procedure were within the nominal levels for tests with 10% of the items showing DIF and higher than expected in a few cases for tests with 20% DIF items. The few cases for the equal ability

Table 6
Mean Percent Detection Rates (Power) for the SIB and MH Procedures for the U_5 and $U_{1.0}$ Ability Distributions Under all Conditions for $\alpha = .05$ and $\alpha = .01$

Ability	10% DIF				20% DIF			
Distribution	SIB		MH		SIB		MH	
and Factor	.05	.01	.05	.01	.05	.01	.05	.01
U_5 Ability Distribution								
Sample Size								
$N_R = 300, N_F = 100$	61	47	58	45	59	43	56	42
$N_R = 300, N_F = 200$	74	62	70	55	74	60	70	57
$N_R = 300, N_F = 300$	82	71	77	67	79	67	72	61
$N_R = 500, N_F = 100$	64	51	62	49	60	48	60	48
$N_R = 500, N_F = 200$	80	69	75	65	80	68	74	62
$N_R = 500, N_F = 300$	86	77	81	72	85	76	79	68
$N_R = 1{,}000, N_F = 100$	67	54	65	51	62	51	61	50
$N_R = 1{,}000, N_F = 200$	84	74	78	64	82	72	76	65
$N_R = 1{,}000, N_F = 300$	89	81	85	77	88	80	82	74
Type of Item								
Low *b*, Medium *a*	91	80	92	83	89	78	91	83
Low *b*, High *a*	96	91	94	86	89	82	94	89
Medium *b*, Low *a*	73	58	69	55	69	57	67	50
Medium *b*, High *a*	95	90	93	90	93	88	93	87
High *b*, Low *a*	51	34	41	24	50	33	37	20
High *b*, Medium *a*	55	38	42	25	54	37	39	21
DIF Effect Size (Area)								
.4	55	38	50	35	52	34	47	33
.6	75	61	70	57	72	57	67	54
.8	85	75	81	71	85	74	79	68
1.0	92	84	89	81	91	84	87	78
$U_{1.0}$ Ability Distribution								
Sample Size								
$N_R = 300, N_F = 100$	66	54	51	45	63	53	52	42
$N_R = 300, N_F = 200$	76	63	63	53	74	65	63	53
$N_R = 300, N_F = 300$	80	69	69	60	78	69	67	58
$N_R = 500, N_F = 100$	68	54	59	48	66	54	56	45
$N_R = 500, N_F = 200$	80	70	67	59	79	69	66	57
$N_R = 500, N_F = 300$	86	78	74	65	84	75	71	62
$N_R = 1{,}000, N_F = 100$	70	58	60	50	68	55	58	48
$N_R = 1{,}000, N_F = 200$	82	73	70	62	80	71	68	59
$N_R = 1{,}000, N_F = 300$	87	79	75	68	87	78	71	64
Type of Item								
Low *b*, Medium *a*	97	91	96	91	96	92	95	90
Low *b*, High *a*	99	97	99	97	99	95	98	94
Medium *b*, Low *a*	77	61	64	46	75	59	59	42
Medium *b*, High *a*	95	90	92	85	95	89	90	81
High *b*, Low *a*	50	33	26	13	46	31	22	10
High *b*, Medium *a*	44	26	19	8	42	27	17	7
DIF Effect Size (Area)								
.4	62	48	50	40	59	46	47	36
.6	74	63	63	54	73	62	61	51
.8	83	73	72	63	81	73	70	61
1.0	89	81	80	72	88	80	77	68

Table 7
Mean False Positive Rates (Type I Error) for the SIB and MH Procedures for the Equal Ability Distribution Under All Conditions for α = .05 and α = .01

	10% DIF				20% DIF			
	SIB		MH		SIB		MH	
Factor	.05	.01	.05	.01	.05	.01	.05	.01
Sample Size								
$N_R = 300, N_F = 100$	.062	.015	.037	.007	.067	.016	.042	.008
$N_R = 300, N_F = 200$	.052	.010	.036	.006	.060	.014	.045	.008
$N_R = 300, N_F = 300$	.054	.013	.042	.008	.058	.013	.045	.010
$N_R = 500, N_F = 100$	.060	.016	.036	.006	.076	.024	.040	.008
$N_R = 500, N_F = 200$	.052	.011	.038	.006	.061	.014	.047	.009
$N_R = 500, N_F = 300$	.056	.011	.042	.006	.066	.016	.054	.011
$N_R = 1{,}000, N_F = 100$	.063	.020	.038	.008	.077	.026	.042	.008
$N_R = 1{,}000, N_F = 200$	.060	.014	.042	.008	.068	.019	.045	.009
$N_R = 1{,}000, N_F = 300$	.055	.011	.042	.008	.064	.015	.051	.010
Type of Item								
Low *b*, Medium *a*	.057	.019	.039	.008	.091	.030	.053	.011
Low *b*, High *a*	.066	.019	.030	.007	.063	.031	.044	.010
Medium *b*, Low *a*	.062	.017	.036	.007	.063	.017	.037	.005
Medium *b*, High *a*	.053	.012	.042	.008	.061	.013	.052	.008
High *b*, Low *a*	.048	.010	.036	.006	.056	.012	.043	.007
High *b*, Medium *a*	.060	.011	.042	.007	.069	.014	.045	.011

distribution were for $N_R = 500$, $N_F = 300$ and $N_R = 1{,}000$, $N_F = 300$ and low *b*, medium *a* and medium *b*, high *a*. The few cases for the $U_{.5}$ ability distribution were $N_R = 500$, $N_F = 300$ and $N_R = 1{,}000$, $N_F = 300$ and medium *b*, low *a*; medium *b*, high *a*; and high *b*, low *a*. For the $U_{1.0}$ ability distribution, the Type I error rates were higher than expected regardless of whether the tests had 10% or 20% DIF items.

For the SIB procedure, the Type I error rates were slightly higher than expected for the equal, $U_{.5}$, and $U_{1.0}$ ability distributions regardless of whether the tests had 10% or 20% of the items as DIF. The Type I error rates also increased as the ability distribution differences increased (i.e., from $U_{.5}$ to $U_{1.0}$) and as the proportion of items showing DIF increased (from 10% to 20%).

Type of item. The type of item did not seem to affect the Type I error rates for either procedure. At $\alpha = .05$ and $\alpha = .01$, the Type I error rates for the MH procedure were the same as the nominal level, with a few exceptions ($U_{1.0}$ with low *b* and high *b* items).

Discussion and Conclusions

Overall, high agreement was found between the SIB and MH procedures in detecting uniform DIF. As expected, the MH and the SIB procedures were affected by sample size. The power of MH and SIB increased as sample size increased, which was not surprising because empirical distributions are expected to approach theoretical distributions as sample size increases. However, the specific purpose of this study was to investigate the effectiveness of these procedures in small sample sizes in which IRT procedures are not feasible. The question investigated was how small a sample size is sufficient for these procedures to be viable methods for detecting uniform DIF. In general, the results showed that, for both procedures, detection rates were affected by reference as well as focal group sample sizes. In particular, the detection rates for the two procedures were affected more by the small size of the focal group than by the larger reference group sample size. For example, for $N_R = 300$, the detection rates for N_F (100, 200, 300) increased from 62% to 84% (an increase of 22%) and for $N_F = 100$, the detection rates for N_R (300, 500, 1,000) increased from 62% to 66% (an increase of 4%) for the equal ability distribution (Table 5). These results also were observed for the $U_{.5}$ and $U_{1.0}$ ability distribution (Table 6). On average, when the focal group sample size increased from 100 to 300, the

Table 8
Mean False Positive Rates (Type I Error) for the SIB and MH Procedures for the $U_{.5}$ and $U_{1.0}$ Ability Distributions Under all Conditions for $\alpha = .05$ and $\alpha = .01$

Ability	10% DIF				20% DIF			
Distribution	SIB		MH		SIB		MH	
and Factor	.05	.01	.05	.01	.05	.01	.05	.01
$U_{.5}$ Ability Distribution								
Sample Size								
$N_R = 300, N_F = 100$	.061	.016	.036	.006	.066	.019	.041	.008
$N_R = 300, N_F = 200$	.058	.013	.038	.007	.058	.014	.048	.009
$N_R = 300, N_F = 300$	.057	.013	.042	.009	.057	.013	.047	.009
$N_R = 500, N_F = 100$	.065	.012	.039	.008	.069	.023	.046	.009
$N_R = 500, N_F = 200$	.055	.016	.038	.007	.062	.016	.049	.010
$N_R = 500, N_F = 300$	.058	.012	.045	.009	.065	.017	.077	.037
$N_R = 1{,}000, N_F = 100$	.067	.019	.042	.009	.074	.024	.042	.008
$N_R = 1{,}000, N_F = 200$	.058	.015	.042	.009	.060	.015	.049	.010
$N_R = 1{,}000, N_F = 300$	.057	.013	.044	.008	.060	.014	.052	.014
Type of Item								
Low *b*, Medium *a*	.061	.014	.040	.008	.059	.016	.041	.012
Low *b*, High *a*	.058	.014	.051	.008	.059	.017	.052	.010
Medium *b*, Low *a*	.060	.017	.044	.008	.058	.016	.054	.017
Medium *b*, High *a*	.067	.013	.040	.005	.071	.019	.061	.015
High *b*, Low *a*	.058	.014	.048	.011	.069	.017	.060	.014
High *b*, Medium *a*	.055	.014	.046	.009	.059	.014	.047	.014
$U_{1.0}$ Ability Distributions								
Sample Size								
$N_R = 300, N_F = 100$	.068	.020	.042	.009	.078	.022	.041	.010
$N_R = 300, N_F = 200$	.072	.020	.048	.010	.086	.025	.047	.011
$N_R = 300, N_F = 300$	.091	.027	.049	.010	.010	.032	.055	.012
$N_R = 500, N_F = 100$	.072	.021	.042	.009	.078	.023	.051	.011
$N_R = 500, N_F = 200$	.082	.021	.050	.011	.090	.025	.056	.011
$N_R = 500, N_F = 300$	.093	.028	.055	.012	.102	.032	.061	.013
$N_R = 1{,}000, N_F = 100$	.078	.023	.045	.010	.080	.026	.048	.010
$N_R = 1{,}000, N_F = 200$	.080	.021	.057	.014	.087	.025	.060	.014
$N_R = 1{,}000, N_F = 300$	.094	.025	.062	.016	.102	.032	.072	.021
Type of Item								
Low *b*, Medium *a*	.068	.031	.056	.013	.078	.031	.061	.010
Low *b*, High *a*	.072	.023	.059	.013	.080	.025	.063	.035
Medium *b*, Low *a*	.070	.021	.031	.006	.067	.013	.034	.010
Medium *b*, High *a*	.067	.020	.050	.011	.073	.020	.059	.015
High *b*, Low *a*	.062	.022	.059	.013	.067	.019	.074	.020
High *b*, Medium *a*	.071	.024	.056	.011	.067	.019	.064	.017

detection rates increased by approximately 20%; when the reference group sample size increased from 300 to 1,000, the corresponding increase was only approximately 10%. These results suggest that varying the sample size and the ratio of reference group to focal group members will have an impact on the performance of MH and SIB procedures for detecting DIF. Overall, a sample size of $N_R = 300$ and $N_F = 300$ was large enough to provide power for the two procedures to detect a reasonable amount of DIF (for area values of .8 and above).

DIF effect size had a significant effect on DIF detection procedures regardless of the size and ratio of the reference and focal groups. In general, for all sample sizes, the detection rates for both procedures steadily increased as the area values increased from .4 to 1.0. Overall, there was an increase of only approximately 10% to 12% in the detection rates when the focal group sample size increased from 100 to 300 with an area value of 1.0 (high

DIF). There was approximately a 26% to 34% increase in detection when the focal group sample size increased from 100 to 300 and the area value was .4 (low DIF). These detection rates were slightly higher for the unequal ability distributions. The results suggest that items that exhibit very small amounts of DIF may go undetected when sample sizes are 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 also support the findings of Rogers (1989) and Rogers & Swaminathan (1993) that the type of item included is a significant factor influencing the detection rates of the DIF detection procedures. Detection rates were highest for highly discriminating items followed by medium and low discriminating items. Detection rates were lowest for high difficulty items ($b \geq 1.5$) followed by items of medium difficulty and low difficulty. Highly difficult items will not be answered correctly by the majority of reference and focal group members. Therefore, highly difficult items may affect only a small number of examinees because only a very few examinees are likely to be found at the extreme ends of the distributions. Fortunately, very difficult items are not very common in standardized achievement tests and hence they may not be a matter of great concern in practice.

The most interesting finding in this study was that the ability distribution differences between the reference and the focal groups did not have an effect on the SIB procedure, although they did have an effect on the MH procedure. This appears to be due to the regression correction used in the SIB procedure. According to Shealy & Stout (1993), the regression correction adjusts the studied subtest scores for the two groups so that they are now estimates of the same latent ability in the case of no DIF, even if group target ability distribution differences exist. The SIB procedure can be very useful when differences in the reference and focal group ability distributions exist in practical settings.

The percentage of items exhibiting DIF did not affect the DIF detection rates to a large extent. This may be due to the two-stage procedure adopted in computing the SIB and MH statistics. Items identified as DIF in the first stage were removed when forming the score groups for computing the DIF statistics in the second stage. Overall, the results showed that the performance of SIB was better in detecting DIF than MH for the unequal ability distributions under most conditions.

The investigation of the Type I error rates indicated that they were within the nominal limits and conservative for the MH procedure. They were slightly higher for the SIB procedure than for the MH procedure for the equal ability distribution. There appeared to be an inflation of Type I error rates for both procedures as the ability distribution differences increased, and the inflation was slightly higher for the SIB procedure. SIB may be preferred because its Type I error rate was marginally higher (1% to 2%) whereas its power was approximately 25% higher.

Although test length was not an issue in this study, test length may affect Type I error rates. In the case of long (and hence more reliable) tests, the inflation in Type I error rates is likely to be small for both procedures. However, for shorter tests, the regression correction used in the SIB procedure improved its performance relative to MH.

Conclusions

The SIB procedure is as powerful as the MH procedure for detecting uniform DIF when ability distributions are the same. SIB has more power than the MH procedure when the reference and focal group ability distributions are unequal. Both procedures are computationally simple, inexpensive, and require little computer time. Both methods can be used interchangeably when the reference and focal groups have equal distributions.

Although the findings here are consistent with other research, several areas merit further investigation. This research and other studies have indicated that both the MH and SIB procedures are to some extent dependent on sample size. There is need for further research to determine the power of these procedures for small sample sizes taking into consideration the ratio of the reference to focal group sample sizes. Although this research suggests that the SIB procedure is more suitable than the MH procedure for unequal ability distributions, which is often the case in practical settings, more research is needed in this area. Future research should concentrate on comparing estimators of DIF effect sizes and their properties.

References

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29,* 67–91.

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

Clauser, B. E. (1993). Factors influencing the performance of the Mantel-Haenszel procedure in identifying differential item functioning. (Doctoral dissertation, University of Massachusetts at Amherst, MA). *Dissertation Abstracts International, 54,* 493.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization 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.

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.

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

Raju, N. S., Bode, R. K., & Larsen, V. S. (1989). An empirical assessment of the Mantel-Haenszel statistic for studying differential item performance. *Applied Measurement in Education, 2,* 1–13.

Rogers, H. J. (1989). A logistic regression procedure for detecting item bias. (Doctoral Dissertation, University of Massachusetts at Amherst, MA). *Dissertation Abstracts International, 50,* 3928.

Rogers, H. J. (1991). *A FORTRAN V program for computing the Mantel-Haenszel DIF statistics.* New York: Columbia University, Teacher's College.

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

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

Shealy, R., & Stout, W. F. (1991, April). *An item response theory model for test bias.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

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

Shealy, R., Stout, W., & Roussos, L. (1991). *SIBTEST user manual* [Computer program manual]. Champaign: University of Illinois, Department of Statistics.

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

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

Author's Address

Send requests for reprints or further information to Pankaja Narayanan, 152 Hills South, School of Education, University of Massachusetts, Amherst MA 01003, U.S.A.