ED 364 557 TM 020 705

AUTHOR	Narayanan, Pankaja; Swaminathan, H.
TITLE	Performance of the Mantel-Haenszel and Simultaneous Item Bias Procedures for Detecting Differential Item Functioning. Laboratory of Psychometric and Evaluative Research Report No. 252.
INSTITUTION	Massachusetts Univ., Amherst. School of Education.
PUB DATE	Apr 93
NOTE	39p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Atlanta, GA, April 13-15, 1993).
PUB TYPE	Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Ability; Comparative Analysis; Computer Simulation; Control Groups; Experimental Groups; *Item Bias; *Nonparametric Statistics; Sample Size; *Statistical Distributions; *Test Items; Test Validity
IDENTIFIERS	Ability Estimates; Asymptotic Distributions; *Mantel Haenszel Procedure; *Simultaneous Item Bias Procedure; Type I Errors

ABSTRACT

The purpose of this study was to compare two non-parametric procedures, the Mantel-Haenszel (MH) procedure and the simultaneous item bias (SIB) procedure, with respect to their Type I error rates and power, and to investigate the conditions under which asymptotic distributional properties of the SIB and MH were obtained. Data were simulated to reflect a variety of conditions; the factors manipulated were sample sizes, ability distributions of the focal and the reference groups, percent of DIF items in the test, types of item and DIF effect sizes. Investigations of the distribution of the SIB and the MH statistics revealed that the SIB statistic had the theoretical asymptotic distributions when the sample sizes of the focal and reference groups exceeded 200, whereas the MH statistic did not have the theoretical asymptotic distributions under any condition. The MH and the SIB procedures were equally powerful in detecting DIF for equal ability distributions, and the SIB procedure was more powerful for unequal ability distributions than the MH procedure. The Type I error rates for the MH statistic were within limits, whereas they were higher for the SIB statistic than those for the MH statistic. Comparisons between the detection rates of the two procedures were made with respect to the various factors manipulated in the study. Suggestions for future research are made. (Contains 16 references and 11 tables.) (Author)

Performance of the Mantel-Haenszel and Simultaneous Item Bias Procedures for Detecting Differential Item Functioning[1,2]

Pankaja Narayanan, H. Swaminathan
University of Massachusetts at Amherst

[1] Laboratory of Psychometric and Evaluative Research Report No. 252. Amherst, MA: University of Massachusetts, School of Education.

[2] Paper presented at the meeting of NCME, Atlanta, April 1993.

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

Pankaja Narayanan, H. Swaminathan
University of Massachusetts at Amherst

Abstract

The purpose of this study was to compare two non-parametric procedures, the Mantel-Haenszel (MH) procedure and the simultaneous item bias (SIB) procedure, with respect to their Type I error rates and power, and to investigate the conditions under which the asymptotic distributional properties of the SIB and MH were obtained.

Data were simulated to reflect a variety of conditions; the factors manipulated were sample sizes, ability distributions of the focal and the reference groups, percent of DIF items in the test, types of item and DIF effect sizes.

Investigations of the distribution of the SIB and the MH statistics revealed that the SIB statistic had the theoretical asymptotic distributions when the sample sizes of the focal and reference groups exceeded 200, whereas the MH statistic did not have the theoretical asymptotic distributions under any condition. The MH and the SIB procedures were equally powerful in detecting DIF for equal ability distributions, and SIB procedure was more powerful for unequal ability distributions than the MH procedure. The Type I error rates for the MH statistic were within limits, whereas they were higher for the SIB statistic than those for the MH statistic. Comparisons between the detection rates of the two procedures were made with respect to the various factors manipulated in the study. Suggestions for future research are made.

Performance of the Mantel-Haenszel and Simultaneous Item Bias Procedures for Detecting Differential Item Functioning[3,4]

Pankaja Narayanan, H. Swaminathan
University of Massachusetts at Amherst

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. The most theoretically sound procedures are based on item response theory (IRT). However, these procedures require large sample sizes, a condition that is often difficult to meet in practice in most DIF studies. Because of this problem, measurement specialists have been actively involved in developing non-parametric methods as alternatives to IRT procedures.

Some of the recently developed methods for detecting DIF are the Mantel-Haenszel (MH) procedure (Holland and Thayer, 1988), the standardization (STD) procedure (Dorans and Kulick, 1986), the simultaneous item bias (SIB) procedure (Shealy and Stout, 1991) and the logistic regression procedure (Swaminathan & Rogers, 1990). Research conducted on the Mantel-Haenszel procedure has shown it to be one of the most effective methods for detecting DIF (Hambleton & Rogers, 1989; Raju, Bode & Larsen, 1989, Mazor, Clauser & Hambleton, 1992). Swaminathan and Rogers (1990) presented the logistic regression procedure and demonstrated that this procedure is as powerful as the Mantel-Haenszel procedure in detecting uniform DIF and more powerful than the Mantel-Haenszel procedure when detecting non-uniform DIF. In fact, the Mantel-Haenszel procedure may be conceptualized as a special case of the logistic regression model where the ability (or test score) is regarded as a discrete variable and where no interaction between the ability variable and

[3] Laboratory of Psychometric and Evaluative Research Report No. 252. Amherst, MA: University of Massachusetts, School of Education.

[4] Paper presented at the meeting of NCME, Atlanta, April 1993.

group membership is allowed. The major advantage of the logistic regression procedure is that, along with its capability to detect non-uniform DIF, it can be expanded to condition on more than one test or subtest scores.

The Mantel-Haenszel and the simultaneous item bias are two theoretically sound procedures that share a common framework. Both procedures are non-parametric, and therefore do not require model calibration (Ackerman, 1992). Both procedures provide tests of significance, computationally simple and inexpensive.

Both procedures most typically use the raw 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 the focal groups. The SIB procedure requires the identification of a "valid" subtest for matching examinees. 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 the focal group examinees. In addition, the SIB procedure can simultaneously evaluate the DIF present in several test items.

A number of studies comparing the Mantel-Haenszel procedure with other popular DIF detection procedures have indicated the Mantel-Haenszel procedure performed well in detecting uniform DIF with considerably lower cost (Hambleton & Rogers 1989; Swaminathan & Rogers, 1990). However, recent research has indicated that under certain circumstances the MH procedure may have a higher Type I error rate than expected (Zwick, 1990). In general, it appears that the MH procedure has 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 three-parameter item response model rather than a one-parameter model. Roussos (1992), using simulated data, showed that the nominal Type I error rates for the SIB procedure is more acceptable than those of the MH procedure in such cases. Ackerman (1992) demonstrated that in the multiple-biased item case, the SIB procedure with its emphasis on the selection of a "valid" subtest for matching the examinees, performed better than the MH procedure with total score used as the matching criterion.

Research Objectives

While considerable research had been carried out on the MH procedure, relatively little research has been conducted on the SIB procedure. The SIB procedure is relatively new, and given the possibility that it may be superior to the MH procedure under certain circumstances, the focus of this study is a detailed investigation of the performance of the SIB procedure.

The main purposes of this study were to compare the Type I error rates and the power of the MH and the SIB procedures to investigate the conditions under which each procedure is optimal for detecting DIF. The logistic regression procedure was not compared with the SIB procedure because this study was confined to the investigation of uniform DIF. A further purpose of the study was to investigate the conditions under which the asymptotic distributions of the MH and the SIB statistics were obtained.

Description of the DIF Statistics

1. The Mantel-Haenszel Procedure

The Mantel-Haenszel procedure (Holland and Thayer in 1988) compares the probabilities of a correct response in the two groups of interest for examinees of the same ability.

In order to calculate the MH statistic, item response data for the reference and the focal group members are arranged into a series of 2 x 2 contingency tables. One such table is constructed for each test item to accommodate group by item response at each score level. In all, a K 2 x 2 contingency tables are constructed where K is the distinct number of score levels formed for the test. For the ith item and a given number correct score j, a 2 x 2 contingency table is constructed in the form shown below:

	Score on studied item		
	1	0	Total
Group			
Reference	A_j	B_j	Nr_j
Focal	C_j	D_j	Nf_j
Total	$M1_j$	$M0_j$	T_j

From the above K 2 x 2 tables, for a given item, the Mantel-Haenszel statistic is computed as follows:

$$\text{MH Chisq} = \frac{(|\Sigma A_j - \Sigma E(A_j)| - .5)^2}{\Sigma \operatorname{Var}(A_j)},$$

where

A_j = the observed number of examinees in the majority group at score level j answering the item correctly;

$$E(A_j) = (Nr_j M1_j)/T_j;$$

and

$$\operatorname{Var}(A_j) = \frac{(Nr_j Nf_j M1_j M0_j)}{((T_j)^2(T_j - 1))}$$

A weighted average of the odds at each of j score levels, where the weight is obtained using the product of the frequencies of right and wrong responses divided by the frequency of responses. It is known as MH Alpha (α) and provides an estimate for DIF effect size. It is given by

$$\text{MH Alpha} = \frac{\Sigma A_j D_j/T_j}{\Sigma B_j C_j/T_j}$$

MH Alpha value can range from 0 to ∞. When MH Alpha value is 1, it indicates that the correct response is equally likely for both the groups. When MH Alpha value is greater than 1, it indicates that the reference group members are more likely to answer the item correctly and vice versa.

The MH Delta statistic introduced by the Educational Testing Service (ETS) is a statistic which is obtained by a non-linear transformation of MH Alpha. A positive value of MH Delta can be interpreted as indicating that the item was easier for the focal group than for the reference group. MH Delta is given by

$$\text{MH Delta} = -(2.35)\ \ln\ (\text{MH Alpha})$$

3. The Simultaneous Item Bias Procedure

The simultaneous item bias (SIB) procedure developed by Shealy & Stout (1991) is based on the multidimensional IRT model. It provides a statistical test to detect DIF present in one or more items on a test simultaneously.

To test whether a set of items on the test is DIF, 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 (they can be one or more items), and a "valid" subtest containing the items that measure the construct that the test is purported to measure. To calculate the SIB statistic, examinee response data on the "valid" subtest scores 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 n+1. The reference and focal group members with the same valid subtest scores are then arranged to form statistic calculation cells such that each statistic calculation cell will correspond to a particular "valid" subtest score. Within each statistic calculation cell, the average proportion right on the "suspect" subtest is calculated for the reference and the focal groups.

Let $\overline{Y}_{Rk}$ and $\overline{Y}_{Fk}$ be the average score in the "suspect" subtest for all examinees in the reference and the focal groups respectively attaining a "valid" subtest score $X = k$ ($k = 0, 1, 2, \ldots\ldots, n$). Since $\overline{Y}_{Rk} - \overline{Y}_{Fk}$ is the difference in performance in the suspect subtest across the two groups among examinees of the same ability, $\overline{Y}_{Rk} - \overline{Y}_{Fk}$ will be expected to be equal to zero if the suspect subtest items are not DIF. The null hypothesis for testing the statistic is stated as

$$H_0: \beta_U = 0 \quad \text{vs.} \quad H_1: \beta_U > 0,$$

where β_U is a parameter denoting the amount of uniform DIF. The test statistic for testing the hypothesis of no DIF is given by

$$B = \hat{\beta}_U / \hat{\sigma}(\hat{\beta}),$$

where $\hat{\beta}_U = \Sigma\, \hat{p}_k\, (\overline{Y}_{Rk} - \overline{Y}_{Fk})$ and $\hat{p}_k$ is the proportion among the focal group examinees attaining $X = k$ on the "valid" subtest and $\hat{\sigma}(\hat{\beta}_U)$ is the estimated standard error of $\hat{\beta}_U$. The null hypothesis of no DIF is rejected at level α if the value of B exceeds the upper $100(1-\alpha)$th percentile point of the standard normal distribution. $\hat{\beta}_U$ is also the statistic used to estimate the amount of DIF. For example, a $\hat{\beta}_U$ value of 0.1 indicates that the average difference in the probabilities of correct response or "studied" subtest score between reference and focal group examinees on similar ability is 0.1.

Research Design

Overview of the Procedure

This research study was conducted on simulated data sets. Examinee response data sets were simulated under a variety of conditions each data set accommodating some of the factors that might have an effect on DIF detection rates. It was decided to do the study on simulated data sets so that it would be possible to specify different amounts of DIF in a set of test items for a variety of conditions and study their effect on items that are a priori known to be differentially functioning.

The study was conducted in two parts. Part one was focused to investigate the distributional properties of the MH and the SIB statistics to determine if the conditions for satisfying the asymptotic distributional properties were obtained. Therefore, in this phase of the study, the research question was whether the MH statistic was distributed as a chi-square distribution with one degree of freedom and the SIB statistic was distributed as a normal distribution with mean zero and standard deviation one. The nominal Type 1 error rates of the MH and SIB statistics were compared to investigate the viability of the two procedures for detecting DIF.

Part two of the study investigated the power of the SIB procedure to determine its potential for detecting the presence of uniform DIF in test items. The performance of the SIB procedure relative to the MH procedure was also examined. It was decided to confine this study to the investigation of uniform DIF because it occurs more commonly than non-uniform DIF.

Description of the Study of the Asymptotic Distributional Properties of the MH and SIB procedures

Since the distributional properties are asymptotic, it can be expected that as the sample size increases, the empirical sampling distribution of the test statistics is more likely to approach the theoretical distribution. Therefore, sample sizes were manipulated to study their effect on the asymptotic distributional properties.

Since in practice the sample sizes of the minority groups may be small, often ranging from 100-300 examinees, three levels of reference group sample sizes (300,500,1000) were crossed with three levels of focal group sample sizes (100,200,300) to give a total of nine tests. Ability values for the two groups were randomly sampled from a normal distribution with mean zero and standard deviation one.

The distributional aspect of the study was conducted using a test length of 45 items. For each of the nine sample sizes, response data for a 45-item test were simulated one each for the reference and the focal groups. The same item parameter values were specified for the reference and the focal groups to represent items in which no DIF were present. To represent a realistic situation, item parameter values for a set of items used to simulate response data were randomly chosen from published tables of item parameter values obtained during an administration of the Graduate Management Admission Test (Kingston, Leary & Wightman, 1988).

Out of the 45 items, the distributional properties of the test statistics were studied for five items with prespecified item parameter values. Different combinations of item parameter values for these five items were chosen to represent various levels of difficulty (b) and discrimination (a). They are: (1) low b (-1.50), low a (0.50); (2) low b (-1.50), high a (1.50); (3) medium b (0.0), medium a (1.0); (4) high b (1.50), low a (0.50);

and (5) high b (1.50), high a (1.50). The c-parameters for these five items were set equal to .20. The five combinations of item characteristics are reported in Table 1.

Since the three-parameter model has been seen to adequately fit many types on data including data with multiple-choice items with four or five options per item, data for the study were simulated for a three-parameter model using the program DATAGEN (Hambleton & Rovinelli, 1973).

The distributions of the test statistics for the SIB and MH procedures across 1000 replications of the data were obtained. To test the asymptotic properties of the MH and SIB statistics, the Kolmogorov-Smirnov (K-S) and the Wilks-Shapiro (W-S) tests were carried out wherever appropriate. The K-S goodness-of-fit test would indicate if the MH statistic has a chi-square distribution with one degree of freedom and if the SIB statistic has a normal distribution with a mean zero and standard deviation one. The Wilks-Shapiro goodness-of-fit test would also indicate if the conditions for the normality of a distribution is satisfied and is therefore, appropriate for the SIB statistic.

Description of the Study of the Power of the MH and SIB Procedures

In this phase of the study, the power of the SIB and MH statistics was studied under a variety of conditions likely to have an impact on DIF. The power of SIB relative to that of MH statistic was also examined.

One factor of interest concerned the size of the sample for the majority and minority groups. Research conducted on the effect of sample sizes on the power of the MH procedure has indicated that DIF detection rıtes increased with increase in sample sizes (Rogers, 1989; Swaminathan & Rogers, 1990; Mazor et al., 1992). In general, when sample size increases, the power of DIF detection procedures will also increase.

A second factor of interest was the ability distribution differences between the two groups. Mazor et al. (1992) have studied the effects of MH procedure when two groups were sampled from equal and unequal distributions. They recommend that, when groups of differing abilities are to be compared, it is probably advisable to use large sample sizes. For the SIB procedure also, it is expected that the detection rates for groups of differing abilities may be different from the detection rates for groups of equal abilities.

A third factor of interest was the proportion of items containing 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.

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 is expected to increase as well. The power of the DIF procedures for different DIF effect sizes were examined to reflect a variety of conditions and compared to determine their capability to detect DIF under these circumstances.

The DIF effect sizes were determined using an IRT framework. In IRT framework, DIF is said to exist if the ICCs for the two groups are not the same. Therefore, the difference between the ICCs for the two groups can be used as a measure of DIF effect size. If the difference between the ICCs is large, then the DIF effect size will also be large and vice versa. Swaminathan and Rogers (1990) used the area between the two ICCs as an operational measure of DIF effect size. They have investigated area values ranging from 0.2 to 0.8. For the purpose of this study, the area between the two ICCs was used as an operational measure of DIF.

The four factors described above were varied to result in nine levels of sample size, three levels of ability distribution differences and two levels of proportion of items containing DIF. Under each condition, four levels of DIF effect size and six types of item were studied. In all 1296 conditions were simulated.

Three reference group sample sizes (300, 500, 1000) were crossed with three focal group sample sizes (100, 200, 300) to produce nine sample sizes. The study was confined to a single test length of 40, a length which is typical of standardized achievement and ability subtests. The impact of the differences in underlying ability distributions was investigated by examining three different conditions. In the first case, ability distributions for the two groups were set to be equal with a mean 0 and standard deviation one. In the second case, the mean was set to 0.0 and -0.5 for the reference and focal groups respectively, with both standard deviations set equal to one. This will be referred to as unequal distribution 1. Distributions that differ by 0.5 standard deviation were specified to simulate the case where there is not a very substantial between group difference. In the third case, the mean was set to 0.0 and -1.0 for the reference and the focal groups respectively, again with both standard deviations set equal to one. Distributions that differ by one standard deviation were chosen to simulate the case where there is a substantial between group difference. To study the effect of the proportion of items containing DIF, tests were simulated with either 10% or 20% of the items containing DIF. It is seen in practice that standardized achievement test usually contain up to about 10% to 15% items as DIF. The 20% proportion of DIF items was included to represent the "worst case scenario".

To simulate items containing DIF, the item param:ter values were chosen similar to Swaminathan & Rogers (1990). Four levels of DIF effect size were chosen equal to the area values .4, .6, .8 and 1.0. Area values in this range would reflect DIF effect sizes that are likely to be found in practice. The item parameter values were chosen so that the areas between the ICCs for the two groups were approximately equal the four DIF effect size values. The areas between the ICCs were chosen by using the formula given by Raju (1988). Since different combinations of discrimination parameters and difficulty parameters can yield a required area, six different combinations of item parameters representing different levels of difficulty and discrimination (low, medium, high) were crossed with four DIF effect sizes to yield 24 DIF items for the study.

Data were generated using the program DATAGEN (Hambleton & Rovinelli, 1973) for a number of tests to investigate the capability of the SIB and MH procedures to identify items that are a priori known to be differentially functioning. Item parameter values were randomly chosen from published item parameter values from an administration of the Graduate Management and Admissions Test (Kingston, Leary & Wightman, 1988). The c-parameters for all the items were set equal to .20.

Six 40-item tests were simulated to contain 10% of the items exhibiting DIF. On each of the six tests, the item parameter values for 36 items were common (same across all the tests). They were also kept common to the reference and focal groups to represent items that were not differentially functioning. The 24 items investigated for DIF were distributed across the six tests. In order to obtain tests containing 10% DIF, the 24 items studied items were distributed equally across the six tests to contain four items in each test. In a similar manner, three 40-item tests containing 20% DIF were simulated. The item parameter values for 32 items were kept common to all the three tests. The 24 studied items were distributed across three 40-item tests to include eight items in each test. Table 1 presents the item parameters values chosen for the distribution and DIF studies.

Insert Table 1 about here

In summary, DIF analyses were carried out for datasets simulated for nine sample sizes, three ability distribution differences, tests containing 10% or 20% DIF items and DIF effect sizes in terms of area between the ICCs for the two groups. In all 1296 conditions were studied. The data were replicated 100 times for each condition.

Results

Study of the Distributional Assumptions of the SIB and MH procedures

The Kolmogorov-Smirnov (K-S) and Wilks-Shapiro (W-S) test results for investigating the distributional properties of the SIB and MH statistics are presented in Tables 2 through 4 respectively.

Insert Tables 2 through 4 about here

The main findings for the SIB statistics (Table 2) are as follows:

1. The means and the standard deviations of most of the 45 empirical distributions closely approximated the mean (0.0) and the standard deviation (1.0) of the theoretical distributions (normal).
2. The K-S goodness-of-fit results show that the theoretical distributions were not obtained for five of 45 conditions. Three of these occurrences were for focal group sample sizes of 100 and one each

for focal group sample sizes of 200 and 300. The reference group sample size for all these occurrences was equal to 300.

3. Four of the above five conditions occurred for items with high difficulty and the remaining for an item with medium difficulty.

4. The W-S goodness-of-fit results (Table 2) confirm the K-S test results, and provide further evidence that the normality assumptions of the SIB test statistic were satisfied for all 45 conditions. The W-S test does not specifically test for conditions of normality with mean zero and standard deviation 1.0.

The main findings of the MH statistic (Table 3) are as follows:

1. The means and the standard deviations of the empirical chi-square distributions of most of the 45 items were lower than the mean (0.0) and the standard deviaticn (1.414) of the theoretical chi-square distributions.

2. The MH statistic was not distributed as a chi-square distribution under any of the conditions studied here.

The Type I error rates for the MH and SIB statistics (Table 4) are as follows:

1. The observed Type I error rates of the SIB statistic were higher than Type I error rates of the MH statistic. At .05 level of statistical significance, Type I errors rates for SIB ranged from about 4.2% to about 7.5%. At .01 level of statistical significance, they ranged from about 0.5% to about 2.5%. The Type I error rates for the MH procedure were within expected limits. At .05 level of statistical significance, Type I error rates varied from 2.5 % to about 5% and at .01 l vel of statistical significance, they were between 0.1% to 1%.

Study of the Power of the SIB and MH Procedures

The analyses in this phase of the study focus on the power of the SIB and MH procedures to detect the 24 studied items presented in Table 1. The DIF detection rates revealed in Tables 5 through 8 are summarized and presented below.

Insert Tables 5 through 8 about here

An analysis of variance (ANOVA) was performed to determine the effects of the five conditions on the performance of SIB and MH procedures. 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. Table 5 shows the ANOVA results for the detection rates across all conditions for the SIB and MH statistics.

The ANOVA results demonstrated a general trend in the results. A review of ANOVA results show that for both SIB and MH procedures, sample size, percent of items containing DIF, type of item and DIF effect size have significant main effects at .05 level of statistical significance.

Several interaction effects were observed for both procedures. These were sample size with ability distribution differences, sample size with type of item, sample size with DIF effect size, ability distribution differences with type of item, ability distribution differences with DIF effect size, and type of item with DIF effect size were all significant. Interestingly, for both procedures, the percent of DIF factor did not have any interaction with the other four factors, namely, sample size, ability distribution differences, type of item and DIF effect size.

Table 6 through 8 present the mean percent of items correctly identified as differentially functioning by the two procedures under all conditions, namely, sample sizes, types of items, DIF effect sizes in terms of area values, and proportion of items containing DIF, for equal and unequal ability (1) and (2) distributions respectively.

The main findings (Tables 6 through 8) for the two procedures are as follows:

Sample Size

1. For equal ability, unequal ability (1) and unequal ability (2) distributions (Table 6 through 8), the detection rates for the two procedures showed a steady increase for increase in the three levels of reference and focal group sample sizes. There was an overall decrease in the detection rates of about 1% to 5% for the two procedures as the proportion of DIF items increased from 10% to 20%. In general, the detection rates for both procedures showed a similar pattern when tests contained 10% DIF and 20% DIF. In most cases, the SIB procedure identified higher percentage of DIF items than the MH procedure for unequal ability distributions.

Type of Item

1. For equal ability distribution (Table 6), the detection rates for the two procedures were highest for highly discriminating, moderate difficulty items followed by low difficulty items. The lowest detection rates were for high difficulty, low discrimination items followed medium discrimination items. In general, as the difficulty level of the items increased, the power of the two DIF procedures decreased. On the other hand, as the discrimination level of the items increased, the power of the two DIF procedures increased. As the percentage of DIF items on the test increased, the detection rates decreased. On the whole, the

detection rates for the two procedures showed a similar pattern for tests containing 10% and 20% DIF.

2. The results for unequal ability (1) and (2) distributions (Table 7 and 8), reveal that for medium difficulty items, the detection rates for the two procedures were comparable with those obtained with equal ability distributions. For low difficulty items, the detection rates for both procedures were better than those obtained with equal ability distributions for tests containing 10% DIF and 20% DIF. The detection rates for high difficulty items were lower for both procedures than those obtained with equal ability distribution.

3. A comparison of the detection rates of the two procedures showed that for medium difficulty, low discrimination items, MH identified about 5% to 7% fewer items for unequal (1) and unequal (2) distributions respectively. In contrast, SIB had similar identification rates for equal ability distributions irrespective of whether tests contained 10% DIF or 20% DIF.

4. For low difficulty items, the detection rates for the two procedures increased by about 8% to 12% for unequal (1) and (2) distributions respectively for both procedures over those obtained for equal ability distribution irrespective of whether tests contained 10% DIF or 20% DIF.

5. The detection rates for high difficulty, low discrimination items reduced by about 7% and 15% for unequal (1) and 8% to 30% for unequal (2) distributions respectively for the SIB and MH procedures.

6. For items of high difficulty, medium discrimination the detection rates for SIB and MH procedures reduced by 10% and 22% and by 22% and 45% for both unequal (1) and (2) distributions respectively.

7. Overall, the SIB procedure was able to identify more items as DIF for unequal ability distributions than the MH procedure. In fact for certain item types, SIB was able to detect about 25% more items as DIF than MH when the ability distributions were unequal.

DIF Effect Sizes

1. For equal as well as unequal (1) and (2) ability distributions (Tables 6 through 8), the detection rates for the two procedures steadily increased for increase in the area values from .4 to 1.0 for each sample size.

The next step in the analyses was to determine the Type I error rates (number of non-DIF items falsely identified as DIF) for the two procedures. Tables 9 through 11 present the mean Type I error rates for the two procedures for equal and two unequal ability distributions respectively.

Insert Tables 9 through 11 about here

The main findings are:

Sample Size

1. Sample size did not seem to affect Type I error rates for both procedures.
2. On the whole, the SIB procedure had higher Type I error rates than the MH procedure.
3. For equal ability distribution (Table 9), at .05 and .01 levels of statistical significance, the mean percent Type I error rates for the MH procedure were within limits for all the sample sizes with a few exceptions. There was little difference in these rates for tests containing 10% or 20% DIF. For the SIB procedure, the mean Type I error

rates ranged up to about 6.3% for tests containing 10% DIF and up to about 7.7% for tests containing 20% DIF.

4. For unequal ability (1) distribution (Table 10), the mean percent Type I error rates for the MH procedure were also within limits for all sample sizes except one case when it was greater than expected, when tests contained 20% DIF. Again, SIB results revealed that the mean percent Type I error rates were those obtained for the MH procedure.

5. For unequal ability (2) distribution (Table 11), the mean percent Type I error rates were inflated for both procedures. These inflations ranged up to about 6.2% for tests containing 10% DIF and about 7.2% for tests containing 20% DIF for the MH procedure. For the SIB procedure these numbers on the whole, ranged up to about 10%.

In Tables 9 through 11, the mean Type I error rates for selected items with different combinations of item parameters are presented to determine if certain types of items were more likely to be incorrectly classified as DIF. In most cases, a pattern emerged much like the results reported for sample sizes with respect to the three ability distributions and percent of DIF items. Again, the type of item did not seem to have an effect on Type I error rates.

Discussion

The results of the first part of the study indicate that for most types of items, the SIB statistic has the expected distribution for reference and focal group for all sample sizes. Items for which the theoretical distributions were not obtained were highly difficult items. The MH statistic appears not to be distributed as a chi-square distribution with one degree freedom for all sample sizes and for all types of items.

The results also suggest that the estimated means and standard deviations of the distributions of the SIB statistics are more acceptable than those of the MH statistics which, are for almost all cases, seem to be underestimated. The Type I error rates for the SIB procedure appear to be somewhat higher than expected, whereas, for the MH statistic, they are within the nominal limits. In one sense it can be argued that the Type I error rate at α level of significance should not exceed

$$\alpha + Z_{\alpha/2} \sqrt{}(\alpha(1-\alpha)/n)$$

In our situation where $\alpha = .05$, $n = 1000$, the Type I error rate should not exceed

$$.05 + (1.96) \sqrt{}[(.05 \text{ X } .95)/1000] = .0635$$

or 6.35%. Using this criterion, it is seen that the Type I error rates for the SIB statistic is within expected limits. Obviously the Type I error rates for the MH statistic are well within the limits. Depending upon the use of the test, the practitioner should consider whether to use the MH, a conservative procedure which yields a few false positives and therefore likely to miss a small percentage of items with DIF, or the SIB procedure, which has somewhat higher detection rates, but also has higher false positive rates.

The main findings of the DIF study indicates that, overall, there is high agreement between the SIB and MH procedures in detecting uniform DIF. As can be expected, the MH as well as the SIB procedure are affected by the sample size. The increase in the power of DIF statistics for increase in sample size is not surprising since the empirical distributions are expected to get closer to the theoretical as sample sizes increase. However, the specific purpose of this study was to investigate the effectiveness of these procedures in samples so small where IRT procedures are not feasible. The question therefore becomes, how small a sample size is sufficient for these procedures to be viable methods for detecting uniform DIF. The results show that detection rates are a function of reference as well as focal group sample sizes for both procedures. Detection rates for the two procedures in this study appear to be more dependent on focal group sample size than reference group sample size. In general, on an average, when the focal group sample sizes increased from 100 to 300, the detection rates increased by about 20% whereas, when the reference group sample sizes increased from 300 to 1000, the corresponding increase was only about 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 (300,300) was seen to be sufficient to provide power for the two procedures to detect reasonable amount of DIF.

These results also suggest that besides sample size, as expected, DIF effect size can have a significant effect on DIF detection procedures irrespective of the size and ratio of reference and focal group members. For all sample sizes, the detection rates both procedures steadily increased as area values increased from .4 to 1.0. Overall, there was an increase of only about 10% to 12% in the detection rates for increase in the focal group sample size from 100 to 300 when the area value was 1.0 (high DIF). There was about 26% to 34% increase in the detection rates for increase in the focal group sample size from 100 to 300 when the area value was .4 (low DIF). These numbers were slightly higher for unequal ability distributions. Practitioners should be aware that items which exhibit very small amounts of DIF may go undetected especially 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) that the type of item included is a significant factor influencing the detection rates of the DIF detection procedures. Detection rates were highest for high discrimination items followed by moderate and low discriminating items. Detection rates were lowest for high difficulty items followed by items of moderate difficulty and low difficulty. Highly difficult items will not be answered correctly by the majority of reference and focal group members. Therefore, most difficult items may affect only a small number of examinees as there are likely to be only a very few number of examinees 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 percentage of items containing 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 computations were removed when forming the score groups for computing the DIF statistics for the second time. Overall, the results show that the performance of SIB was higher than MH for unequal ability distributions in most conditions.

As revealed in the distribution study, the power results also indicate that, the Type I error rates were within acceptable limits and conservative for the MH procedure. They were somewhat higher for the SIB procedure than those obtained for the MH procedure for equal ability distributions. There appeared to be inflation of Type I error rates for both procedures as the ability distribution differences increased, the inflation was higher for the SIB procedure. Again, SIB procedure should be used to depending upon how much Type I errors are tolerable in practice.

Conclusions

A comparison between simultaneous item bias procedure and the Mantel-Haenszel procedures indicate that the simultaneous item bias procedure is as powerful as the Mantel-Haenszel procedure for detecting uniform differential item functioning and has more power than the Mantel-Haenszel procedure when the reference and focal group ability distributions are unequal. Both procedures are computationally simple, inexpensive and require little computer time. Both methods are therefore interchangeable and can be used under appropriate situations.

Although the results in this research are consistent with the findings of several other recent research, several areas merit further investigation. The asymptotic distributional properties of both statistics need to be examined for unequal ability distributions also to determine the viability of both procedures under these conditions. This research and other studies indicate that both Mantel-Haenszel and simultaneous item bias 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 simultaneous item bias procedure is more suitable than the Mantel-Haenszel procedure for unequal ability distributions, more research is needed in this area. Future research should concentrate on comparing estimators of DIF effect sizes and their properties. Some of these issues will be addressed in a future study by the authors of this paper.

References

Ackerman, T. 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 meeting of NCME, San Francisco, CA.

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.

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 Mantel-Haenszel procedure. In H. Wainer and H. I. Braun (Eds.), *Test validity* (129-145). Hillsdale, NJ: Lawrence Erlbaum Associates.

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.

Rogers, H. L. (1989). *A Logistic Regression Procedure for detecting item bias*. Unpublished Doctoral Dissertation, University of Massachusetts at Amherst, MA.

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.

Roussos, L. A. (1992). *A comparison of inflation of nominal Type I error rates for bias/DIF detection using SIBTEST and the Mantel-Haenszel*. Technical Report, University of Illinois, Urbana-Champaign.

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

Shealy, R. & Stout, W. F. (in press). 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.

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.

Table 1. Item Parameters Used to Generate Items for the Distribution and DIF Studies

Item	Type of Item	DIF effect Size	Ref. a1	Foc. a2	Ref. b1	Foc. b2
Distribution Study						
1.	Low a Low b		.50	.50	-1.50	-1.50
2.	High a Low b		1.50	1.50	-1.50	-1.50
3.	Medium a Medium b		1.00	1.00	0.00	0.00
4.	Low a High b		.50	.50	1.50	1.50
5.	High a High b		1.50	1.50	1.50	1.50
DIF Study						
1	Low a Medium b	.4	.50	.50	-0.26	0.26
2		.6	.50	.50	-0.39	0.39
3		.8	.50	.50	-0.51	0.51
4		1.0	.50	.50	-0.64	0.64
5	Low a High b	.4	.50	.50	1.28	1.80
6		.6	.50	.50	1.14	1.92
7		.8	.50	.50	1.01	2.04
8		1.0	.50	.50	0.88	2.16
9	Medium a Low b	.4	.90	.90	-1.80	-1.28
10		.6	.90	.90	-1.92	-1.14
11		.8	.90	.90	-2.04	-1.01
12		1.0	.90	.90	-2.16	-0.88
13	Medium a High b	.4	.90	.90	1.28	1.80
14		.6	.90	.90	1.14	1.92
15		.8	.90	.90	1.01	1.24
16		1.0	.90	.90	0.88	2.16
17	High a Low b	.4	1.25	1.25	-1.80	-1.28
18		.6	1.25	1.25	-1.92	-1.14
19		.8	1.25	1.25	-2.04	-1.01
20		1.0	1.25	1.25	-2.16	-0.88
21	High a Medium b	.4	1.25	1.25	-0.26	0.26
22		.6	1.25	1.25	-0.39	0.39
23		.8	1.25	1.25	-0.51	0.51
24		1.0	1.25	1.25	-0.64	0.64

Table 2. Kolmogorov-Smirnov and Wilks-Shapiro Test Results for Testing the Distributional Assumptions of the SIB Statistic

Sample	Size	Item	Mean	S.D.	K-S	p	W-S	p
Ref	Foc				Abs. Diff.			
300	100	1	0.01	1.05	.033	.21	.989	.90
		2	0.04	1.02	.030	.29	.985	.31
		3	0.00	.96	.051	.01*	.991	.97
		4	-0.04	1.00	.060	.00*	.986	.53
		5	0.00	.94	.055	.00*	.984	.15
300	200	1	0.02	.97	.026	.53	.987	.64
		2	0.00	1.04	.024	.59	.989	.99
		3	0.02	1.02	.026	.49	.992	.99
		4	0.05	1.01	.022	.70	.986	.40
		5	0.06	1.01	.047	.02*	.990	.95
300	300	1	0.00	.98	.026	.52	.985	.90
		2	-0.04	1.02	.032	.26	.988	.75
		3	-0.05	1.03	.029	.38	.987	.49
		4	0.09	1.00	.057	.00*	.987	.51
		5	-0.04	.99	.033	.22	.988	.71
500	100	1	-0.05	1.06	.029	.36	.984	.10
		2	-0.07	1.08	.035	.17	.983	.06
		3	-0.01	1.03	.016	.96	.990	.95
		4	-0.01	1.04	.028	.40	.985	.21
		5	-0.01	1.03	.027	.46	.986	.34
500	200	1	0.00	.97	.025	.56	.986	.43
		2	-0.01	1.01	.026	.53	.987	.55
		3	0.00	1.01	.018	.89	.988	.75
		4	-0.07	1.01	.037	.13	.989	.25
		5	-0.04	.98	.028	.40	.987	.52
500	300	1	-0.03	1.01	.025	.55	.990	.92
		2	-0.02	.99	.027	.46	.988	.69
		3	0.03	.99	.021	.79	.991	.97
		4	0.05	.99	.036	.15	.986	.30
		5	0.00	1.01	.030	.33	.984	.10
1000	100	1	-0.03	1.10	.026	.51	.982	.06
		2	-0.12	1.08	.043	.06	.987	.66
		3	0.02	1.02	.021	.72	.987	.54
		4	-0.06	1.07	.035	.18	.986	.47
		5	0.05	1.00	.026	.53	.983	.07
1000	200	1	-0.03	0.97	.029	.37	.987	.60
		2	-0.05	1.03	.026	.50	.988	.65
		3	0.03	1.05	.032	.16	.986	.34
		4	-0.02	1.02	.020	.82	.990	.91
		5	0.04	.99	.032	.24	.989	.85
1000	300	1	0.03	1.01	.030	.29	.987	.49
		2	-0.02	.98	.026	.51	.986	.32
		3	0.00	1.03	.023	.63	.988	.77
		4	0.01	1.00	.025	.53	.991	.98
		5	0.02	1.03	.030	.34	.989	.89

Table 3. Kolmogorov-Smirnov Test Results for Testing the Distributional Assumptions of the MH Statistic

Sample Size		Item	Mean	S.D.	K-S	p
Ref	Foc				Abs. Diff.	
300	100	1	.75	1.21	.128	.00
		2	.66	1.12	.200	.00
		3	.78	1.19	.126	.00
		4	.77	1.16	.110	.00
		5	.79	1.22	.111	.00
300	200	1	.77	1.18	.121	.00
		2	.73	1.28	.174	.00
		3	.89	1.38	.098	.00
		4	.90	1.38	.114	.00
		5	.81	1.34	.112	.00
300	300	1	.84	1.31	.116	.00
		2	.78	1.33	.135	.00
		3	.81	1.24	.113	.00
		4	.87	1.33	.113	.00
		5	.82	1.21	.107	.00
500	100	1	.79	1.17	.113	.00
		2	.62	0.98	.216	.00
		3	.88	1.37	.112	.00
		4	.79	1.23	.122	.00
		5	.82	1.34	.109	.00
500	200	1	.75	1.10	.110	.00
		2	.79	1.17	.130	.00
		3	.82	1.24	.104	.00
		4	.85	1.23	.092	.00
		5	.85	1.31	.104	.00
500	300	1	.85	1.31	.114	.00
		2	.81	1.28	.135	.00
		3	.86	1.28	.113	.00
		4	.90	1.41	.086	.00
		5	.85	1.29	.092	.00
1000	100	1	.85	1.31	.106	.00
		2	.67	1.06	.167	.00
		3	.78	1.17	.106	.00
		4	.84	1.31	.114	.00
		5	.75	1.19	.119	.00
1000	200	1	.81	1.26	.107	.00
		2	.74	1.16	.141	.00
		3	.86	1.26	.101	.00
		4	.88	1.28	.113	.00
		5	.81	1.31	.138	.00
1000	300	1	.88	1.30	.114	.00
		2	.75	1.17	.140	.00
		3	.90	1.38	.104	.00
		4	.91	1.32	.114	.00
		5	.87	1.35	.112	.00

Table 4. Percent of Replications for which Non-DIF Items were Falsely Identified as DIF for the SIB and MH Procedures

			SIB Statistic		MH Statistic	
Sample Size		Item	α=.05	α=.01	α=.05	α=.01
Ref	Foc		(%)	(%)	(%)	(%)
300	100	1	5.8	1.2	3.3	0.7
		2	6.8	2.5	3.1	0.5
		3	5.5	0.8	3.8	0.9
		4	4.8	1.1	3.5	0.5
		5	5.6	1.9	3.6	0.8
300	200	1	4.7	0.9	3.3	0.7
		2	5.7	1.9	3.5	0.9
		3	5.0	1.1	3.8	0.6
		4	5.1	1.6	3.9	0.9
		5	4.9	0.8	3.6	0.4
300	300	1	6.0	1.6	3.5	0.8
		2	5.8	1.7	3.8	0.8
		3	3.4	0.8	3.6	0.6
		4	4.6	0.8	3.3	0.2
		5	4.6	0.5	3.3	0.6
500	100	1	6.0	2.1	3.3	0.4
		2	7.0	3.0	1.6	0.2
		3	7.5	2.5	4.2	0.9
		4	5.9	1.7	3.0	0.9
		5	6.2	2.0	3.3	0.9
500	200	1	4.2	0.6	2.9	0.4
		2	6.5	1.5	3.3	0.1
		3	4.8	1.0	3.2	0.6
		4	5.5	1.1	4.3	0.4
		5	6.2	1.1	3.3	0.7
500	300	1	5.2	1.3	4.1	1.0
		2	5.2	1.6	3.7	1.0
		3	4.8	1.1	3.8	0.7
		4	6.0	1.6	5.4	1.0
		5	5.5	1.0	3.8	0.9
1000	100	1	7.3	2.4	4.1	1.0
		2	8.6	2.8	2.9	0.3
		3	6.9	1.5	2.6	0.6
		4	6.9	2.6	4.1	1.0
		5	4.9	1.9	3.7	0.3
1000	200	1	5.4	1.1	3.7	0.6
		2	6.4	1.3	2.9	0.3
		3	5.5	1.8	4.0	0.7
		4	5.3	0.9	4.3	0.8
		5	4.7	1.4	3.4	0.9
1000	300	1	5.6	0.7	4.9	0.4
		2	4.6	0.7	2.7	0.5
		3	4.8	1.5	4.1	1.1
		4	5.9	1.1	3.5	0.9
		5	4.5	1.3	4.3	0.9

Table 5. Analysis of Variance of the Effects of all Factors on the Performance of the Simultaneous Item Bias and Mantel-Haenszel Procedures on Differential Item Functioning

	SIB		MH	
Factor	F	p	F	p
Main Effects				
Sample Size	273.60	.000*	209.95	.000*
Ability Distribution	0.65	.520	260.50	.000*
Percent DIF	31.95	.000*	39.49	.000*
Type of Item	1737.39	.000*	2878.89	.000*
DIF Effect Size	1958.71	.000*	1857.50	.000*
Interaction Effects				
Sample Size X Ability Distribution	3.32	.000*	2.83	.000*
Sample Size X Percent of DIF	.30	.992	.22	.986
Sample Size X Type of Item	10.27	.000*	6.84	.000*
Sample SIze X DIF Effect Size	5.63	.000*	3.03	.000*
Ability Distribution X Percent of DIF	.03	.975	.02	.980
Ability Distribution X Type of Item	76.64	.000*	184.12	.000*
Ability DIstribution X DIF Effect Size	42.58	.000*	38.52	.000*
Percent of DIF X Type of Item	.99	.423	1.73	.124
Percent of DIF X DIF Effect SIze	1.93	.123	0.69	.560
Type of Item X DIF Effect Size	69.62	.000*	73.73	.000*

Table 6. Mean Percent Detection Rates for the SIB and MH Procedures for Equal Ability Distribution Under all Conditions

Factor		10% DIF				20% DIF			
		SIB		MH		SIB		MH	
Sample Size		α=.05	α=.01	α=.05	α=.01	α=.05	α=.01	α=.05	α=.01
		(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)
Ref	Foc								
300	100	62	45	62	47	60	43	60	44
300	200	78	67	77	64	74	61	72	58
300	300	84	72	82	70	81	70	78	66
500	100	62	46	64	50	61	48	63	48
500	200	81	69	80	69	79	68	77	65
500	300	87	79	87	76	83	72	84	76
1000	100	66	49	69	55	65	48	67	52
1000	200	84	73	85	74	82	70	82	71
1000	300	90	82	90	82	88	78	88	79
Type of Item									
Low a	Medium b	73	59	73	59	70	55	70	54
Low a	High b	58	40	56	36	55	37	51	34
Medium a	Low b	85	71	85	56	81	68	83	72
Medium a	High b	66	52	64	48	66	50	63	45
High a	Low b	88	76	89	81	85	75	86	79
High a	Medium b	95	90	95	93	93	87	94	88
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

Table 7. Mean Percent Detection Rates for the SIB and MH Procedures for Unequal Ability (1) Distribution Under all Conditions

		10% DIF				20% DIF			
Factor		SIB		MH		SIB		MH	
		$\alpha=.05$	$\alpha=.01$	$\alpha=.05$	$\alpha=.01$	$\alpha=.05$	$\alpha=.01$	$\alpha=.05$	$\alpha=.01$
Sample Size		(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)
Ref	Foc								
300	100	61	47	58	45	59	43	56	42
300	200	74	62	70	55	74	60	70	57
300	300	82	71	77	67	79	67	72	61
500	100	64	51	62	49	60	48	60	48
500	200	80	69	75	65	80	68	74	62
500	300	86	77	81	72	85	76	79	68
1000	100	67	54	65	51	62	51	61	50
1000	200	84	74	78	64	82	72	76	65
1000	300	89	81	85	77	88	80	82	74
Type of Item									
Low a	Medium b	73	58	69	55	69	57	67	50
Low a	High b	51	34	41	24	50	33	37	20
Medium a	Low b	91	80	92	83	89	78	91	83
Medium a	High b	55	38	42	25	54	37	39	21
High a	Low b	96	91	94	86	89	82	94	89
High a	Medium b	95	90	93	90	93	88	93	87
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

Table 8. Mean Percent Detection Rates for the SIB and MH Procedures for Unequal Ability (2) Distribution Under all Conditions

		10% DIF				20% DIF			
Factor		SIB		MH		SIB		MH	
		α=.05	α=.01	α=.05	α=.01	α=.05	α=.01	α=.05	α=.01
Sample Size		(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)
Ref	Foc								
300	100	66	54	51	45	63	53	52	42
300	200	76	63	63	53	74	65	63	53
300	300	80	69	69	60	78	69	67	58
500	100	68	54	59	48	66	54	56	45
500	200	80	70	67	59	79	69	66	57
500	300	86	78	74	65	84	75	71	62
1000	100	70	58	60	50	68	55	58	48
1000	200	82	73	70	62	80	71	68	59
1000	300	87	79	75	68	87	78	71	64
Type of Item									
Low a	Medium b	77	61	64	46	75	59	59	42
Low a	High b	50	33	26	13	46	31	22	10
Medium a	Low b	97	91	96	91	96	92	95	90
Medium a	High b	44	26	19	8	42	27	17	7
High a	Low b	99	97	99	97	99	95	98	94
High a	Medium b	95	90	92	85	95	89	90	81
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 9. Mean Percent False Positive (Type I Error) Rates for the SIB and MH Procedures for Equal Ability Distribution Under all Conditions

		10% DIF				20% DIF			
Factor		SIB		MH		SIB		MH	
		α=.05	α=.01	α=.05	α=.01	α=.05	α=.01	α=.05	α=.01
Sample Size		(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)
Ref	Foc								
300	100	6.2	1.5	3.7	0.7	6.7	1.6	4.2	0.8
300	200	5.2	1.0	3.6	0.6	6.0	1.4	4.5	0.8
300	300	5.4	1.3	4.2	0.8	5.8	1.3	4.5	1.0
500	100	6.0	1.6	3.6	0.6	7.6	2.4	4.0	0.8
500	200	5.2	1.1	3.8	0.6	6.1	1.4	4.7	0.9
500	300	5.6	1.1	4.2	0.6	6.6	1.6	5.4	1.1
1000	100	6.3	2.0	3.8	0.8	7.7	2.6	4.2	0.8
1000	200	6.0	1.4	4.2	0.8	6.8	1.9	4.5	0.9
1000	300	5.5	1.1	4.2	0.8	6.4	1.5	5.1	1.0
Type of Item									
Low a	Medium b	6.2	1.7	3.6	0.7	6.3	1.7	3.7	0.5
Low a	High b	4.8	1.0	3.6	0.6	5.6	1.2	4.3	0.7
Medium a	Low b	5.7	1.9	3.9	0.8	9.1	3.0	5.3	1.1
Medium a	High b	6.0	1.1	4.2	0.7	6.9	1.4	4.5	1.1
High a	Low b	6.6	1.9	3.0	0.7	6.3	3.1	4.4	1.0
High a	Medium b	5.3	1.2	4.2	0.8	6.1	1.3	5.2	0.8

Table 10. Mean Percent False Positive (Type I Error) Rates for the SIB and MH Procedures for Unequal Ability (1) Distribution Under all Conditions

		10% DIF				20% DIF			
Factor		SIB		MH		SIB		MH	
		α=.05	α=.01	α=.05	α=.01	α=.05	α=.01	α=.05	α=.01
Sample Size		(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)
Ref	Foc								
300	100	6.1	1.6	3.6	0.6	6.6	1.9	4.1	0.8
300	200	5.8	1.3	3.8	0.7	5.8	1.4	4.8	0.9
300	300	5.7	1.3	4.2	0.9	5.7	1.3	4.7	0.9
500	100	6.5	1.2	3.9	0.8	6.9	2.3	4.6	0.9
500	200	5.5	1.6	3.8	0.7	6.2	1.6	4.9	1.0
500	300	5.8	1.2	4.5	0.9	6.5	1.7	7.7	3.7
1000	100	6.7	1.9	4.2	0.9	7.4	2.4	4.2	0.8
1000	200	5.8	1.5	4.2	0.9	6.0	1.5	4.9	1.0
1000	300	5.7	1.3	4.4	0.8	6.0	1.4	5.2	1.4
Type of Item									
Low a	Medium b	6.0	1.7	4.4	0.8	5.8	1.6	5.4	1.7
Low a	High b	5.8	1.4	4.8	1.1	6.9	1.7	6.0	1.4
Medium a	Low b	6.1	1.4	4.0	0.8	5.9	1.6	4.1	1.2
Medium a	High b	5.5	1.4	4.6	0.9	5.9	1.4	4.7	1.4
High a	Low b	5.8	1.4	5.1	0.8	5.9	1.7	5.2	1.0
High a	Medium b	6.7	1.3	4.0	0.5	7.1	1.9	6.1	1.5

Table 11. Mean Percent False Positive (Type I Error) Rates for the SIB and MH Procedures for Unequal Ability (2) Distribution Under all Conditions

		10% DIF				20% DIF			
Factor		SIB		MH		SIB		MH	
		α=.05	α=.01	α=.05	α=.01	α=.05	α=.01	α=.05	α=.01
Sample Size		(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)
Ref	Foc								
300	100	6.8	2.0	4.2	0.9	7.8	2.2	4.1	1.0
300	200	7.2	2.0	4.8	1.0	8.6	2.5	4.7	1.1
300	300	9.1	2.7	4.9	1.0	10.0	3.2	5.5	1.2
500	100	7.2	2.1	4.2	0.9	7.8	2.3	5.1	1.1
500	200	8.2	2.1	5.0	1.1	9.0	2.5	5.6	1.1
500	300	9.3	2.8	5.5	1.2	10.2	3.2	6.1	1.3
1000	100	7.8	2.3	4.5	1.0	8.0	2.6	4.8	1.0
1000	200	8.0	2.1	5.7	1.4	8.7	2.5	6.0	1.4
1000	300	9.4	2.5	6.2	1.6	10.2	3.2	7.2	2.1
Type of Item									
Low a	Medium b	7.0	2.1	3.1	0.6	6.7	1.3	3.4	1.0
Low a	High b	6.2	2.2	5.9	1.3	6.7	1.9	7.4	2.0
Medium a	Low b	6.8	3.1	5.6	1.3	7.8	3.1	6.1	1.0
Medium a	High b	7.1	2.4	5.6	1.1	6.7	1.9	6.4	1.7
High a	Low b	7.2	2.3	5.9	1.3	8.0	2.5	6.3	3.5
High a	Medium b	6.7	2.0	5.0	1.1	7.3	2.0	5.9	1.5