

Running head: EFFECTS OF LARGE DIF

**Performance of Mantel-Haenszel, Simultaneous Item Bias Test, and  
Logistic Regression When the Proportion of DIF Items is Large\*\***

Mark J. Gierl  
Centre for Research in Applied Measurement and Evaluation  
University of Alberta

Michael G. Jodoin  
University of Massachusetts, Amherst

Terry A. Ackerman  
University of North Carolina, Greensboro

Paper Presented at the Annual Meeting of the  
American Educational Research Association (AERA)

New Orleans, Louisiana, USA  
April 24-27, 2000

---

\*\* This paper can also be downloaded from the Centre for Research in Applied Measurement and Evaluation (CRAME) website: <http://www.education.ualberta.ca/educ/psych/crame/>

## Abstract

Differential item functioning (DIF) analyses are used to identify items that operate differently between two groups. Three popular DIF methods—Mantel-Haenszel (MH), Simultaneous Item Bias Test (SIBTEST) and logistic regression (LR)—match examinees on a conditioning variable to ensure that the two groups are comparable to one another. However, in some testing situations like test translation and adaptation, the proportion of DIF items can be large. In such cases, it is not well understood how a large number of DIF items will affect the Type I error rates or the power for these three popular methods. This important problem is addressed in the current study. Three variables were manipulated in a simulation study: The amount of DIF on a 40-item test (20, 40, and 60% of the items on the test had either moderate or large DIF), sample size (250, 500, and 1000 examinees in each group), and ability distribution differences between groups (equal and unequal). Two key results were reported. First, excessive numbers of DIF items in the conditioning variable (i.e., up to 60% of the items) did not adversely affect the Type I error rates for MH, SIBTEST, or LR. In most conditions, Type I error rates were below the nominal alpha level of 0.05 even with small samples of examinees. Consequently, all three methods provide good Type I error protection when the proportion of DIF items is large. Second, power differed markedly across the three methods. SIBTEST was the most powerful method. Moreover, the power of SIBTEST increased substantially in both the equal and unequal ability distributions when the sample was increased from 250 to 500 examinees per group compared to MH and LR. We conclude that DIF analyses can be conducted, with accuracy, when the proportion of DIF items is large. These results are reassuring for researchers and practitioners alike, especially those who work in the area of test translation and adaptation, since DIF rates on translated tests can be large. MH, SIBTEST, and LR provide adequate Type I error protection, even when 60% of the items contain DIF, but SIBTEST is the most powerful of the three methods.

### **Performance of Mantel-Haenszel, Simultaneous Item Bias Test, and Logistic Regression When the Proportion of DIF Items is Large**

Differential item functioning (DIF) is present when examinees from different groups have a different probability or likelihood of answering an item correctly, after conditioning on ability. Conditioning on ability is a critical step because it ensures that examinees are matched on a common measure before they are compared. To produce accurate results, the conditioning variable must also provide a valid measure of the construct of interest for both groups (Ackerman & Evans, 1994; Clauser, Nungester, Mazor, & Ripkey, 1996). Although many DIF methods are available, a relatively small number of these methods are "preferred" based on their theoretical and empirical strengths (Clauser & Mazor, 1998). Three of the preferred methods frequently used to detect items with DIF are Mantel-Haenszel (MH), Simultaneous Item Bias Test (SIBTEST), and logistic regression (LR).

Increasingly, DIF analyses are being used to identify items that are not comparable between language groups on translated and adapted tests (e.g., Allalouf, Hambleton, & Sireci, 1999; Budgell, Raju, & Quartetti, 1995; Gierl & Khaliq, 2000; Hambleton, 1994). Researchers who study the psychometric characteristics of translated tests have noted an important trend: The amount of DIF on some translated tests is unusually large. For example, Gierl, Rogers, and Klinger (1999) reported that 26 of 50 items (52%) on a Canadian Grade 6 Social Studies achievement test translated from English to French contained DIF. Ercikan (1999) found that 58 out of 140 science items (41%) on the Third International Mathematics and Science Study (TIMSS) contained DIF when the Canadian English and French examinees were compared. Allalouf et al. (1999) noted that 42 out of 125 verbal items (34%) contained DIF on the Israeli Psychometric Entrance Test when Hebrew and Russian examinees were compared. These outcomes have raised questions and concerns about the validity of the conditioning variable in some translation DIF studies.

The selection of a valid, defensible conditioning variable is critical for achieving accurate results. When large numbers of DIF items are found, as in the studies just cited, the accuracy and appropriateness of the conditioning variable becomes questionable. In this situation, some researchers suggest a two-step purification procedure where the first step is used to flag DIF items, these items are then removed from the conditioning variable, and the second step uses the purified (i.e., DIF-free) conditioning variable to flag the DIF items (e.g., Allalouf et al., 1999; Camilli & Shepard, 1994; Dorans & Holland, 1993; Holland & Thayer, 1988; Lord, 1980). Purification is

justifiable because DIF items may degrade ability estimation. This degradation may, in turn, adversely affect DIF detection because the conditional methods use the ability estimate as the matching variable. However, the two-step procedure may be inefficient and unnecessary, especially since no one has documented the effects of DIF on the matching variable when the three conditional methods used in this study are applied to tests with a large proportion of DIF items. This represents an important omission in the literature on differential item functioning. Shealy and Stout (1993) allude to one potential outcome when they reported that the Type I error rate was only marginally affected by a large proportion of DIF items in the matching variable using the SIBTEST procedure (in one condition of their simulation study, 23% of the items in the matching variable contained DIF items) but, due to space limitations, the details from this condition were omitted from their manuscript. Thus, the purpose of the present study is to compare Type I error rates and power for three popular conditional and widely-used methods—Mantel-Haenszel (MH), Simultaneous Item Bias Test (SIBTEST), and logistic regression (LR)—when the proportion of DIF items is large.

#### Overview of Three Conditional DIF Methods

##### Mantel-Haenszel

Mantel-Haenszel (MH) is a nonparametric statistical approach for identifying DIF (Mantel & Haenszel, 1959; Holland & Thayer, 1988). The MH statistic is distributed as a chi-square test with one degree of freedom. The null hypothesis specifies that there is no relation between group membership and test performance on one item after controlling for ability. MH is used to estimate the constant odds ratio that yields a measure of effect size for evaluating the amount of DIF that is present. MH is computed by matching examinees in each group on total test score and then forming a  $K$  2-by-2 contingency tables for each item, where  $K$  is the score level on the matching variable which is usually total test score. At each score level  $j$ , a 2-by-2 contingency table is created for each item  $i$ , as shown in Figure 1.

-----  
 Insert Figure 1 about here  
 -----

With reference to Figure 1, the MH chi-square test is calculated as follows:

$$c_{MH}^2 = \frac{\{|\sum_j [A_j - E(A_j)]| - 0.5\}^2}{\sum_j \text{var}(A_j)},$$

where,

$$E(A_j) = \frac{N_{r_j} T_{1j}}{T_j}$$

and,

$$\text{var}(A_j) = \frac{N_{r_j} N_{f_j} T_{1j} T_{0j}}{T_j^2 (T_j - 1)}.$$

The estimate of the constant odds ratio is:

$$a_{MH} = \frac{\sum_j A_j D_j / T_j}{\sum_j B_j C_j / T_j}.$$

$a_{MH}$  is the ratio of the odds that a reference group examinee (i.e., group serving as the standard of comparison) will get the item correct compared to the odds for a matched focal group examinee (i.e., group believed to be disadvantaged on the test).  $a_{MH}$  is often transformed to  $\Delta_{MH}$  to enhance the interpretability of the result using the formula,

$$\Delta_{MH} = -(2.35) \ln(a_{MH}).$$

Research at the Educational Testing Service has resulted in proposed  $\Delta_{MH}$  values for classifying DIF as negligible, moderate, or large (Zieky, 1993, p. 342; Zwick & Ercikan, 1989). Roussos and Stout (1996) modified these guidelines to aid in the interpretation of DIF as follows:

- Negligible or A-level DIF: Null hypothesis is retained or null hypothesis is rejected and  $|\Delta_{MH}| < 1$ ,
- Moderate or B-level DIF: : Null hypothesis is rejected and  $1 \leq |\Delta_{MH}| < 1.5$ ,
- Large or C-level DIF: Null hypothesis is rejected and  $|\Delta_{MH}| \geq 1.5$ .

These ratings are used to classify DIF items in the present study.

### Simultaneous Item Bias Test

The Simultaneous Item Bias Test (SIBTEST) is an alternative nonparametric statistical method for detecting DIF proposed by Shealy and Stout (1993). SIBTEST is intended to calculate the size of DIF using a multidimensional perspective. With this method, the complete latent space is viewed as multidimensional,  $(\Theta, \eta)$ , where  $\Theta$  is the unidimensional construct of interest or the target ability and  $\eta$  is the extraneous or nuisance abilities.

The statistical hypothesis tested by SIBTEST is:

$$H_0: B(T) = P_R(T) - P_F(T) = 0$$

vs.

$$H_1: B(T) = P_R(T) - P_F(T) \neq 0,$$

where  $B(T)$  is the difference in probability of a correct response on the studied item for examinees in the reference and focal groups matched on true score;  $P_R(T)$  is the probability of a correct response on the studied item for examinees in the Reference group with true score  $T$ ; and  $P_F(T)$  is the probability of a correct response on the studied item for examinees in the Focal group with true score  $T$ . In other words  $B(T)$ , the parameter representing the amount of unidimensional DIF when a single test item is evaluated, is 0 when there is no DIF and nonzero when DIF is present. With the SIBTEST approach, items on the test are divided into two subsets, the suspect subtest and the matching subtest. The suspect subtest contains the biased item and the matching subtest contains the rest of the items. For each matching subtest score,  $k$ , the corresponding subtest true score for the Reference and Focal groups is estimated using linear regression. The estimated true scores are then adjusted using a regression correction technique to ensure the estimated true score is comparable for the examinees in the Reference and Focal groups on the matching subtest. In the final step,  $B(T)$  is estimated using  $\hat{B}$  which is the weighted sum of the differences between the proportion-correct true scores on the studied item for examinees in the two groups across all score levels.

More succinctly, the test statistic for assessing the null hypothesis in SIBTEST is:

$$\hat{B} = \frac{\hat{B}_U}{\hat{s} \sqrt{\hat{B}_U}},$$

where,

$$\hat{B}_U = \sum_k \hat{p}_k \left( \bar{Y}_{R_k}^* - \bar{Y}_{F_k}^* \right),$$

and,

$$\hat{s} \sqrt{\hat{B}_U} = \left[ \sum_k \hat{p}_k \left( \frac{1}{J_{R_k}} \hat{s}^2(Y|k, R) + \frac{1}{J_{F_k}} \hat{s}^2(Y|k, F) \right) \right]^{1/2}.$$

In the formula for the SIBTEST test statistic,  $\hat{p}_k$  is the proportion of examinees in the Focal group

obtaining  $X = k$  on the valid subtest;  $\bar{Y}_{R_k}^*$  and  $\bar{Y}_{F_k}^*$  are the adjusted means for examinees in

subgroup  $k$  (adjusted using a regression correction procedure outlined in Shealy & Stout, 1993);

$\hat{s}^2(Y|k, g)$  is the sample variance for examinees in the studied subtest for group  $g$  (i.e., the

Reference and Focal group) with a total score of  $k$  on the valid subtest; and  $J_{g_k}$  is the sample size

for group  $g$  with a total score of  $k$  on the valid subtest.

Like MH, SIBTEST yields an overall statistical test as well as a measure of the effect size for each

item ( $\hat{B}$  is an estimate of the amount of DIF). Roussos and Stout (1996, p. 220) proposed the

following  $\hat{B}$  values for classifying DIF as negligible, moderate, and large:

- Negligible or A-level DIF: Null hypothesis is rejected and  $|\hat{B}| < 0.059$ ,
- Moderate or B-level DIF: Null hypothesis is rejected and  $0.059 \leq |\hat{B}| < 0.088$ ,
- Large or C-level DIF: Null hypothesis is rejected and  $|\hat{B}| \geq 0.088$ .

These guidelines are adopted in the current study to identify DIF items.

SIBTEST differs from MH in a number of ways. First, SIBTEST uses a regression estimate of the true score instead of an observed score as the matching variable. As a result, examinees are matched on an estimated latent ability score rather than an observed score. Second, SIBTEST can be used to evaluate DIF in two or more items simultaneously in the analysis. This feature allows the developer to assess DIF more effectively in testlets or item bundles on a test (Douglas, Roussos, & Stout, 1996). Although MH may be considered the 'gold standard' in DIF detection (Roussos & Stout, 1996), researchers have demonstrated that SIBTEST has superior statistical characteristics compared to MH for detecting uniform DIF (the distinction between uniform and non-uniform DIF is described in the next section) (Narayanan & Swaminathan, 1994; Roussos & Stout, 1996; Shealy & Stout, 1993).

### Logistic Regression

A third statistical approach, which is parametric unlike MH and SIBTEST, commonly used to identify DIF is logistic regression (LR; Swaminathan & Rogers, 1990). Logistic regression can detect uniform and non-uniform DIF which provides a distinction between this approach with MH since the later method was only designed to detect uniform DIF. Uniform DIF exists when there is no interaction between ability level and group membership. That is, the probability of answering an item correctly is greater for one group consistently over all ability levels. Non-uniform DIF occurs when there is an interaction between ability level and group membership. In this case, the difference in the probabilities of a correct response for the two groups is not the same at all levels of ability. Simulation studies have been conducted to demonstrate that LR is more powerful than MH at detecting non-uniform DIF (Rogers & Swaminathan, 1993).

With LR, the presence of DIF is determined by testing the improvement in model fit that occurs when a term for group membership and a term for the interaction between test score and group membership are successively added to the logistic regression model. A chi-square test is then used to evaluate the presence of uniform and non-uniform DIF on the item of interest by successively testing each term included in the model. The general model for logistic regression takes the form:

$$P(u = 1) = \frac{e^Z}{1 + e^Z},$$

where  $u$  is the score on the studied item. Performance on the studied item is first conditioned on total test score. In this step,  $z = \mathbf{b}_0 + \mathbf{b}_1X$ , where  $X$  is the test score (Model 1). This serves as the baseline model. The presence of uniform DIF is then tested by examining the improvement in chi-square model fit associated with adding a term for group membership ( $G$ ) against the baseline model. That is, Model 2 (i.e.,  $z = \mathbf{b}_0 + \mathbf{b}_1X + \mathbf{b}_2G$ ) is subtracted from Model 1. The presence of non-uniform DIF is tested by examining the improvement in chi-square model fit associated with adding a term for group membership ( $G$ ) and a term for the interaction between test score and group membership ( $XG$ ) against model 2. In other words, Model 3 (i.e.,  $z = \mathbf{b}_0 + \mathbf{b}_1X + \mathbf{b}_2G + \mathbf{b}_3XG$ ) is subtracted from Model 2. Non-uniform DIF can be tested with LR regardless of the outcome from the uniform DIF test because each model contains different terms.

Jodoin and Gierl (2000) recently evaluated the use of an effect size measure, called  $R^2\Delta$  (Zumbo, 1999; Zumbo & Thomas, 1996), with logistic regression in an attempt to reduce the inflated Type I errors often associated with this approach (Narayanan & Swaminathan, 1996; Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990).  $R^2\Delta$  is given as:

$$R^2\Delta = R^2_2 - R^2_1,$$

where  $R^2_2$  and  $R^2_1$  are the sums of the products of the standardized regression coefficient for each explanatory variable and the correlation between the response and each explanatory variable (i.e.,

$\sum_1^j \mathbf{b}_j r_j$  for  $j$  explanatory variables) for the augmented and baseline models, respectively. By

manipulating the percentage of items containing DIF, sample size, and ability distribution differences and studying the outcomes in the context of a simulation study, Jodoin and Gierl found that  $R^2\Delta$  could, in fact, be used to reduce Type I errors. They also presented new guidelines for interpreting the results from this approach by comparing  $R^2\Delta$  with  $\hat{B}$ , the effect size measure used with SIBTEST. They proposed the following guidelines:

- Negligible or A-level DIF: Null hypothesis is retained or null hypothesis is rejected and  $R^2\Delta < 0.035$ ,
- Moderate or B-level DIF: Null hypothesis is rejected and  $0.035 \leq R^2\Delta < 0.070$ ,

- Large or C-level DIF: Null hypothesis is rejected and  $R^2\Delta \geq 0.070$ .

These guidelines are applicable to both uniform and non-uniform DIF, and they used in the present study to identify DIF items.

#### Method

A simulation study was conducted to compare Type I error rates and power for the MH, SIBTEST, and LR methods when the proportion of DIF items was large. Three variables were manipulated: The amount of DIF on a 40-item test (20, 40, and 60% of the items on the test had either moderate or large DIF), sample size (250, 500, and 1000 examinees in each group), and ability distribution differences between groups (equal and unequal).

#### Amount of DIF

The main variable in this study was the proportion of items with DIF. The proportion of DIF items was manipulated to be 20, 40, and 60% of the items on the 40-item test (i.e., 8, 16, and 24 items with DIF, respectively). For each of the three amounts of DIF, the items had either a 75/25 split with 75% of the items having moderate or B-level DIF and 25% of the items having large or C-level DIF or a 50/50 split with 50% moderate DIF and 50% large DIF to produce, in total, six conditions for this variable. We expected that the moderate DIF items would be more difficult to flag but this situation likely represents a practical reality since most DIF is classified as moderate in large-scale testing programs (e.g., Linn, 1993). In all six conditions, DIF was balanced so half the items favored the reference group and half favored the focal group. This outcome is intended to model testing situations where group differences are unsystematic. Group differences on translated tests often follow this pattern because translation errors tend to be random across the items on a test.

#### Sample Size

Sample size is another important variable because it directly affects power. Therefore, three sample sizes were considered: 250, 500, and 1000 examinees in both the reference and focal groups. For each analysis, the sample sizes were the same for each group.

#### Ability Distribution Differences

Ability distribution differences, the third variable in the simulation study, were also manipulated. Although several simulation studies have demonstrated adherence to nominal Type I error rates with ability differences as large as one standard deviation between reference and focal groups (Narayanan

& Swaminathan, 1994; Rogers & Swaminathan, 1993), ability differences are common when disparate groups, as one might find in translation DIF studies, are compared. Thus, it was included. Two levels were considered. In the equal ability distribution condition, both the reference and focal group abilities were normally distributed with a mean of 0.0 and a standard deviation of 1.0. In the unequal ability distribution condition, the ability distribution for the focal group was normal but with a mean of -0.5 and a standard deviation 1.0.

Taken together, 36 conditions were studied. A 6 (Percent of DIF Items) X 3 (Sample Size) X 2 (Ability Distribution) fully-crossed design was used. Each condition was replicated 100 times.

#### Data Generation and Analysis

The three-parameter logistic item response model was used to generate the examinee response vectors for both non-DIF and DIF items. The non-DIF items included in each test were selected from a translated achievement test used in Canada. The same item parameters were used for both the reference and focal groups resulting in unbiased test items. The item parameters are shown in Table 1. The first 32, 24, and 16 items served as the non-DIF items for the 20, 40, and 60% DIF conditions, respectively.

DIF was created by altering the item parameters between the reference and focal groups. The DIF items were also designed to be realistic. Each test contained DIF items with realistic characteristics such as one might find in a large-scale achievement test. Items with a range of discrimination (i.e.,  $a$ -values of .75, 1.00, and 1.25) and difficulty (i.e.,  $b$ -values of -1.00, .00, and 1.00) were crossed to create the DIF items (see Tables 2 and 3). The  $c$ -value was fixed at .20 for each item. In this study only uniform DIF was modeled. This decision reflects the fact that non-uniform DIF, while theoretically plausible, is rare in practice (Camilli & Shepard, 1994, p. 66; Gierl & McEwen, 1998; Ronald K. Hambleton, personal communication, January 28, 2000; Mazor, Clauser, & Hambleton, 1994). This finding has also been reported in the test translation literature by researchers using methods that are sensitive to non-uniform DIF<sup>1</sup>. DIF effect sizes, based on the area between item response functions (Raju, 1988), were set to 0.4 to reflect moderate or B-level DIF and 0.8 to reflect large or C-level DIF. For the 75/25 split, the following combinations were used to create the DIF items for a 40-item test: The first six items in Table 2 (moderate DIF) and the first two items in Table 3 (large DIF) for the 20% DIF condition; the first 12 items in Table 2 and the first four items in Table 3

for the 40% DIF condition; and the first 18 items in Table 2 and the first six items in Table 3 for the 60% DIF condition. For the 50/50 split, the following item combinations were used: The first four items in Tables 2 and 3 for the 20% DIF condition; the first eight items in Tables 2 and 3 for the 40% DIF condition; and the first 12 items in Tables 2 and 3 for the 60% DIF condition. DIF was balanced so that half the items favored the focal group.

Items were flagged using the conventions for A, B, and C-level DIF, as previously described, since this classification approach is commonly used in practice. Items meeting the A-level criteria were considered non-DIF items whereas items meeting the B- or C-level criteria were considered DIF items. This interpretation seems justified since B- and C-level items are typically scrutinized for potential bias in test reviews (e.g., Zieky, 1993). An alpha level of .05 was used for all hypothesis testing.

-----  
 Insert Tables 1, 2, and 3 about here  
 -----

## Results

The results are presented in Tables 4 to 9. Each table contains the Type I error and power for MH, SIBTEST, and LR as a function of ability distribution differences (equal or unequal) and sample size (250, 500, and 1000). The results for the 75/25 split are presented first, followed by the 50/50 split.

### Type I Error and Power for 75% Moderate DIF and 25% Large DIF

20% DIF. Table 4 contains the results for the condition with 20% DIF items in the conditioning variable. With equal ability distributions between reference and focal groups, Type I error rates remained at or below the nominal alpha level of .05 for all three DIF methods even when the number of examinees was only 250 per group. LR exceeded the nominal level in the 250/250 condition (.0544) while both MH and SIBTEST remained below the nominal level (.0372 and .0472, respectively). Type I error rates were below the nominal level in the 500/500 and 1000/1000 conditions for all three methods. Power, on the other hand, differed markedly between the three methods. SIBTEST was the most powerful method for all three sample sizes. SIBTEST also showed a large increase in power from the 250/250 to 500/500 condition (.7763 to .9438) compared to MH

(.7488 to .8438) and LR (.7288 to .8275). Power exceeded .90 for both SIBTEST and LR in the 1000/1000 condition but not for MH.

-----  
 Insert Table 4 about here  
 -----

Similar results were found with unequal ability distributions. Although LR, again, produced the highest Type I error rate in the 250/250 condition (.0497), all three methods produced acceptable Type I error rates remaining below the nominal alpha level across the three sample sizes. With an unequal ability distribution between the reference and focal group, power decreased. Despite this effect, SIBTEST was still the most powerful method across all three sample sizes, and it showed a large increase in power from the 250/250 to 500/500 condition (.7363 to .9100) compared to MH (.7075 to .8125) and LR (.6988 to .8138). Power exceeded .90 for both SIBTEST and LR in the 1000/1000 condition but remain comparatively low for MH (.8163).

40% DIF. Table 5 contains the results for the condition with 40% DIF items in the conditioning variable. Despite the increase in the proportion of DIF items, similar results to the 20% condition were found. With equal ability distributions, Type I error rates remained below the nominal alpha level of .05 for all three DIF methods. Power differed by methods with SIBTEST having the strongest results across all three sample sizes. SIBTEST showed a large increase in power from the 250/250 to 500/500 condition (.7706 to .9444) compared to MH (.7413 to .8675) and LR (.7163 to .8388). Power exceeded .90 for both SIBTEST and LR in the 1000/1000 condition but not for MH. With unequal ability distributions, all three methods produced acceptable Type I error rates and SIBTEST was the most powerful method across the three sample sizes. SIBTEST showed a large increase in power from the 250/250 to 500/500 condition (.7519 to .9275) compared to MH (.7281 to .8283) and LR (.7138 to .8188). Power exceeded .90 for both SIBTEST and LR in the 1000/1000 condition but not for MH.

-----  
 Insert Table 5 about here  
 -----

60% DIF. Table 6 contains the results for the condition with 60% DIF items in the conditioning variable. With equal ability distributions, Type I error rates, once again, remained below the nominal level for all three DIF methods even when 60% of the items on the test had DIF. Power differed by method with SIBTEST having the strongest results across all three sample sizes. SIBTEST showed a large increase in power from the 250/250 to 500/500 condition (.7750 to .9254) compared to MH (.7500 to .8175) and LR (.7188 to .8121). Power exceeded .90 for both SIBTEST and LR in the 1000/1000 condition but not for MH. With unequal ability distributions, Type I error rates remained close to the nominal alpha level. SIBTEST exceeded the nominal level in the 250/250 condition (.0600) while both MH and LR remained below the nominal level (.0438 for both methods). Type I error rates were below the nominal level in the 500/500 and 1000/1000 conditions for all three methods. For the power analysis, SIBTEST showed a large increase in power from the 250/250 to 500/500 condition (.7213 to .9004) compared to MH (.7004 to .7775) and LR (.6908 to .8008) and the power exceeded .90 for only the SIBTEST method—the power for MH and LR in the 1000/1000 condition was comparatively less at .7813 and .8725, respectively.

-----  
 Insert Table 6 about here  
 -----

#### Type I Error and Power for 50% Moderate DIF and 50% Large DIF

20% DIF. Table 7 contains the results when 20% of the items in the conditioning variable had DIF. With equal ability distributions, Type I error rates were below the nominal alpha level of .05 for all three DIF methods. The three methods were also more powerful in the 50/50 split compared to the 75/25 split because this condition contained more large DIF items which were easier to flag for the three methods. Nevertheless, SIBTEST remained the most powerful method across all three sample sizes, and it showed a large increase in power from the 250/250 to 500/500 condition (.8650 to .9588) compared to MH (.8500 to .8600) and LR (.8250 to .8663). Power exceeded .90 for both SIBTEST and LR in the 1000/1000 condition but not for MH.

With unequal ability distributions, similar results were found. All three methods produced acceptable Type I error rates remaining below the nominal alpha level across sample size. With 250 examinees per group, LR was the most powerful method (.7563 compared to .7475 for MH and .7525

for SIBTEST). However, the power for SIBTEST increased dramatically from 250 to 500 examinees per group (.7525 to .9350) compared to MH (.7475 to .8088) and LR (.7563 to .8425). Power exceeded .90 for both SIBTEST and LR in the 1000/1000 condition but remain comparatively low for MH (.8350).

-----  
 Insert Table 7 about here  
 -----

40% DIF. Table 8 contains the results when 40% of the items in the conditioning variable had DIF. With equal ability distributions, Type I error rates remained below the nominal alpha level of .05 for all three DIF methods. Power differed by method with SIBTEST having the strongest results in all three sample sizes and with SIBTEST showing a large increase in power from the 250/250 condition to 500/500 condition (.8575 to .9569) compared to MH (.8413 to .8831) and LR (.8163 to .8744). Power exceeded .90 for all three methods in the 1000/1000 condition. With unequal ability distributions, both SIBTEST and LR produced Type I errors that exceeded the nominal alpha level of .05 (.0571 and .0517, respectively). Type I error rates either dropped or remained below the nominal level in the 500/500 and 1000/1000 conditions for all three methods. As before, SIBTEST was the most powerful method across the sample sizes and SIBTEST showed a large increase in power from the 250/250 to 500/500 conditions (.8156 to .9394) compared to MH (.7875 to .8394) and LR (.7819 to .8600). Power exceeded .90 for both SIBTEST and LR in the 1000/1000 condition but not for MH.

-----  
 Insert Table 8 about here  
 -----

60% DIF. Table 9 contains the results when 60% of the items in the conditioning variable had DIF. With equal ability distributions, Type I error rates remained below the nominal level for all three DIF methods. Power differed by method with SIBTEST having the strongest results across all three sample sizes and with SIBTEST showing a large increase in power from the 250/250 condition to 500/500 condition (.8483 to .9654). However, unlike the previous results, the increase in power was noticeable for MH (.8267 to .9113) and LR (.8104 to .9008) as well. Power exceeded .90 for all three methods with 1000 examinees per group. With unequal ability distributions, Type I error rates

remained below the nominal alpha level for all methods. SIBTEST was the most powerful method, and it showed a large increase in power from the 250/250 to 500/500 conditions (.8354 to .9508) compared to MH (.8063 to .8904) and LR (.7929 to .8938). Power exceeded .90 for SIBTEST and LR but not for MH (although MH was very close at .8963).

-----  
 Insert Table 9 about here  
 -----

### Conclusions and Discussion

The purpose of the present study was to compare Type I error rates and power for the Mantel-Haenszel (MH), Simultaneous Item Bias Test (SIBTEST), and logistic regression (LR) methods when the proportion of DIF items is large. This study is relevant to researchers and practitioners alike, especially those who study the psychometric characteristics of translated tests or use results from translated tests since it has been found that the amount of DIF on these exams can be large. When this outcome occurs, conditioning on any score derived from the individual item responses may be inappropriate. This research was conducted to document the effects that large proportions of DIF item may exert on the conditioning variable for the three methods used in this study. To-date, no one has addressed this problem in the DIF literature. Three features of this study are noteworthy: (a) three popular as well as theoretically and empirically defensible DIF detection methods were compared, (b) for each of the three methods, DIF items were flagged using a statistical test and an effect size measure to classify items as A-, B-, or C-level, as is currently the practice in many large-scale testing programs, and (c) the DIF methods were evaluated under conditions one might find with translated and adapted tests.

Two key outcomes were reported. First, Type I error rates, for the most part, remained at or below the nominal alpha level of .05 even with samples as small as 250 examinees per group despite the large amount of DIF in the conditioning variable. The largest Type I error was found for SIBTEST in the 250/250 condition with unequal ability distributions. It was .0600 (see Table 6). Overall, MH, SIBTEST, and LR had excellent Type I error rates under all conditions.

Second, SIBTEST was clearly the most powerful DIF detection method. While the power for all three methods was moderate (i.e., in the range .70 to .80) for the 250/250 conditions with the 75/25

split, a sizable increase in power occurred for the 500/500 conditions using SIBTEST compared to the other methods. For example, when power was averaged over the equal and unequal abilities for the results in Tables 4 to 6, SIBTEST was noticeably more powerful than either MH and LR (0.9229 vs. 0.8186 and 0.8163, respectively) in the 500/500 condition. In the 1000/1000 condition, the average power of SIBTEST increased to 0.9598 while it remained in the moderate range for MH (0.8370) and moderate to high range for LR (0.8997).

With a 50/50 split in the DIF items, power increased for all three methods simply because items with large DIF were easier to flag. Nevertheless, SIBTEST continued to show a sizable increase in power from 250 to 500 examinees per group compared to MH and LR. When power was averaged over the equal and unequal abilities for the results in Tables 7 to 9, SIBTEST was more powerful than either MH or LR (0.9529 vs. 0.8766 and 0.8801, respectively) in the 500/500 condition. In the 1000/1000 condition, the overall power increased to 0.9776 for SIBTEST compared to .8963 for MH and .9351 for LR.

This research has three implications for practice. These implications are related to purification, sample size, and DIF method of choice. Dorans and Holland (1993) provide this description of the two-step purification procedure:

Because all tests are imperfect, they in fact may contain some items that do have DIF.

Otherwise, the DIF analysis would be a meaningless exercise. In an attempt to ensure that the matching criterion is in fact DIF-free, DIF analyses at ETS occur in two steps. The first step is called the criterion refinement or purification step. Here, items on the matching variable are analyzed for DIF, and any items that exhibit sizeable DIF are removed regardless of the sign of the DIF. Then, this refined criterion is used for another DIF analysis of the same items and any other items that were not included in the criterion refinement step. (pp. 60-61)

In other words, two-step purification is an attempt to refine the matching variable to produce more accurate DIF results—and two-step purification may have this desired effect. However, the small number of studies to-date are inconclusive. Miller and Oshima (1992), for example, found that the two-step purification procedure did not have a substantial impact on MH DIF detection when the proportion of DIF items was small (i.e., 5 or 10%) but did when the proportion of DIF items was larger

(i.e., 20 or 40%). When the proportion was large, MH produced fewer Type I errors when purification was used. Kwak, Davison, and Davenport (1999) reported that MH without purification would provide adequate Type I error protection for uniform DIF based on their simulation study where 10% of the items contained DIF. But they added that if non-uniform DIF was a concern, then MH may not be the best method. In both studies, MH DIF detection was based on statistical criteria without an associated effect size measure.

Despite these outcomes, most practitioners would agree that purification is costly and time consuming. The results from the current study suggest that this procedure may be unnecessary, especially when large samples are used. The three methods included in this study had low Type I error rates and moderate or high power even when 60% of the items in the conditioning variable contained DIF (see Tables 6 and 9). However, more research is needed to understand this issue. In the present study, DIF was balanced so half the items favored the focal group to reflect the outcomes often found in translation DIF studies given that translation errors tend to be random across the items on a test. These results may not apply to conditions with systematic bias where items consistently favor one group. In addition, only uniform DIF was studied since non-uniform DIF is quite rare in practice. As a next step, the authors of the current manuscript are studying these problems by comparing the Type I error rates and power for MH, SIBTEST, and LR using both purified and unpurified conditioning variables under diverse testing conditions that contain both systematic bias and non-uniform DIF.

The results from this study also have implications for sample size selection. Roussos and Stout (1996) reported that MH and SIBTEST adhered to the nominal level of significance with samples as small as 100 examinees per group. In the current study, Type I error rates also remained close to the nominal alpha level of .05 with samples of 250 examinees per group for MH, SIBTEST, and LR despite the large amount of DIF—up to 60%—in the conditioning variable. However, with samples of 250 examinees per group, the power of the three methods was moderate ranging from approximately .70 to .87 across all conditions. Power increased noticeably with 500 examinees per group, especially for SIBTEST. Thus, researchers and practitioners should include at least 500 examinees per group when using the methods described in this study under similar testing conditions.

Finally, SIBTEST performed extremely well in this simulation study. It produced results at or below the nominal alpha level in all conditions and it was the most powerful DIF detection method. While the power for all three methods was moderate for the 250/250 conditions, SIBTEST showed a sizable increase in power for the 500/500 conditions when compared to MH and LR. Moreover, the overall power for SIBTEST was high at .9689 with 1000 examinees per group when averaged across all conditions in this study. It appears to be a very effective method when large numbers of DIF items are expected.

## References

- Ackerman, T. A., & Evans, J. A. (1994). The influence of conditioning scores in performing DIF analyses. Applied Psychological Measurement, 18, 329-342.
- Allalouf, A., Hambleton, R., & Sireci, S. (1999). Identifying the causes of translation DIF on verbal items. Journal of Educational Measurement, 36, 185-198.
- Budgell, G. R., Raju, N. S., & Quartetti, D. A. (1995). Analysis of differential item functioning in translated assessment instruments. Applied Psychological Measurement, 19, 309-321.
- Camilli, G., & Shepard, L. A. (1994). Methods for identifying biased test items. Newbury Park, CA: Sage.
- Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. Educational Measurement: Issues and Practice, 17, 31-44.
- Clauser, B. E., Nungester, R. J., Mazor, K., & Ripkey, D. (1996). A comparison of alternative matching strategies for DIF detection in tests that are multidimensional. Journal of Educational Measurement, 33, 202-214.
- Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and Standardization. In P. W. Holland & H. Wainer (Eds.) Differential item functioning (pp. 35-66). Hillsdale, NJ: Erlbaum.
- Douglas, J., Roussos, L., and Stout, W. (1996). Item bundle DIF hypothesis testing: Identifying suspect bundles and assessing their DIF. Journal of Educational Measurement, 33, 465-484.
- Erickson, K. (1999, April). Translation DIF on TIMSS. Paper presented at the annual meeting of the National Council on Measurement in Education, Montréal, Quebec, Canada.
- Gierl, M. J., & Khaliq, S. N. (2000, April). Identifying sources of differential item and bundle functioning on translated tests. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.
- Gierl, M. J., & McEwen, N. (1998, May). Differential item functioning on the Alberta Education Social Studies 30 diploma examination. Paper presented at the annual meeting of the Canadian Society for the Study of Education, Ottawa, ON, Canada.

Gierl, M. J., Rogers, W. T., & Klinger, D. (1999, April). Consistency between statistical procedures and content reviews for identifying translation DIF. Paper presented at the annual meeting of the National Council on Measurement in Education, Montréal, Quebec, Canada.

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. European Journal of Psychological Assessment, *10*, 229-244.

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

Jodoin, M. G., & Gierl, M. J. (2000, April). Reducing type I error using an effect size measure with the logistic regression procedure for DIF detection. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Kwak, N., Davison, M. L., & Davenport, E. C. (1999). A comparative study of purification procedures for detecting differential item functioning. Manuscript submitted for publication.

Linn, R. L. (1993). The use of differential item functioning statistics: A discussion of current practice and future implications. In P. W. Holland & H. Wainer (Eds.) Differential item functioning (pp. 349-366). Hillsdale, NJ: Erlbaum.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute, *22*, 719-748.

Mazor, K. M., Clauser, R. E., & Hambleton, R. K. (1994). Identification of nonuniform differential item functioning using a variation of the Mantel-Haenszel procedure. Educational and Psychological Measurement, *54*, 284-291.

Miller, M. D., & Oshima, T. C. (1992). Effect of sample size, number of biased items, and magnitude of bias on a two-stage item bias estimation method. Applied Psychological Measurement, *16*, 381-388.

Narayanan, P., & Swaminathan, H. (1994). Performance of Mantel-Haenszel and simultaneous item bias procedure for detecting differential item functioning. Applied Psychological Measurement, *18*, 315-328.

- Narayanan, P., & Swaminathan, H. (1996). Identification of items that show nonuniform DIF. Applied Psychological Measurement, 20, 257-274.
- Raju, N. S. (1988). The area between two item characteristic curves. Psychometrika, 53, 495-502.
- 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. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type I error performance. Journal of Educational Measurement, 33, 215-230.
- Shealy, R., & Stout, W. F. (1993). A model-based standardization approach that separates true bias/DIF from group differences and detects test bias/DIF as well as item bias/DIF. Psychometrika, 58, 159-194.
- Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. Journal of Educational Measurement, 27, 361-370.
- Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.) Differential item functioning (pp. 337-347). Hillsdale, NJ: Erlbaum.
- Zumbo, B. D. (1999). A handbook on the theory and methods of differential item functioning: Logistic regression modeling as a unitary framework for binary and Likert-type item scores. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.
- Zumbo, B. D., & Thomas, D. R. (1996, October). A measure of DIF effect size using logistic regression procedures. Paper presented at the National Board of Medical Examiners, Philadelphia, PA.
- Zwick, R., & Ercikan, K. (1989). Analysis of differential item functioning in the NAEP history assessment. Journal of Educational Measurement, 26, 55-66.

Author Note

Please address all correspondence to Mark J. Gierl, Centre for Research in Applied Measurement and Evaluation, 6-110 Education Centre North, Faculty of Education, University of Alberta, Edmonton, Alberta, Canada, T6G 2G5. Email: [mark.gierl@ualberta.ca](mailto:mark.gierl@ualberta.ca).

This research was supported with funds awarded to the first author from the Social Sciences and Humanities Research Council of Canada (SSHRC). We would like to thank W. Todd Rogers for his comments on an earlier version of this manuscript.

## Footnote

<sup>1</sup>For example, Gierl, Rogers, and Klinger (1999), in their study of Grade 6 Mathematics and Social Studies achievement test items translated from English to French, found only three non-uniform DIF items from a set of 99 items. Similarly, Ercikan (1999), in her comparison of the English and French Canadian examinees on the science items from the Third International Mathematics and Science Study (TIMSS), flagged only three non-uniform DIF items from a set of 140 items. On the mathematics items, not one displayed non-uniform DIF. LR was used for DIF detection in both studies.

Table 1

Item Parameters for the Negligible or A-Level DIF Items

Item	A	B	C
1	0.71	-2.44	0.25
2	0.91	-1.75	0.24
3	0.97	-1.68	0.22
4	0.82	-1.46	0.24
5	0.60	-1.36	0.29
6	0.76	-1.27	0.25
7	0.59	-1.18	0.28
8	0.81	-1.12	0.17
9	0.59	-1.04	0.22
10	0.76	-0.99	0.32
11	0.89	-0.87	0.19
12	0.89	-0.83	0.34
13	0.61	-0.76	0.21
14	0.87	-0.63	0.33
15	1.15	-0.57	0.20
16	1.04	-0.49	0.16
17	0.66	-0.40	0.24
18	0.76	-0.38	0.20
19	0.77	-0.34	0.31
20	0.81	-0.28	0.20
21	0.96	-0.23	0.30
22	1.21	-0.20	0.12
23	0.62	0.06	0.21
24	0.80	0.15	0.33
25	1.11	0.17	0.24
26	0.87	0.19	0.16
27	0.59	0.24	0.27
28	0.99	0.26	0.12
29	0.66	0.40	0.30
30	0.63	0.52	0.21
31	0.73	0.74	0.13
32	1.32	2.15	0.19

Note. A, B, and C correspond to the discrimination, difficulty, and psuedo-guessing parameters in the 3PL IRT model.

Table 2

Item Parameters for the Moderate or B-Level DIF Items

Item	$A_R$	$B_R$	$C_R$	$A_F$	$B_F$	$C_F$
1	1.00	0.25	0.20	1.00	-0.25	0.20
2	1.00	1.25	0.20	1.00	0.75	0.20
3	1.00	-0.25	0.20	1.00	0.25	0.20
4	1.00	0.75	0.20	1.00	1.25	0.20
5	1.00	-0.75	0.20	1.00	-1.25	0.20
6	1.00	-1.25	0.20	1.00	-0.75	0.20
7	0.75	0.25	0.20	0.75	-0.25	0.20
8	0.75	-0.25	0.20	0.75	0.25	0.20
9	0.75	-0.75	0.20	0.75	-1.25	0.20
10	1.25	0.25	0.20	1.25	-0.25	0.20
11	0.75	-1.25	0.20	0.75	-0.75	0.20
12	1.25	-0.25	0.20	1.25	0.25	0.20
13	1.25	-0.75	0.20	1.25	-1.25	0.20
14	0.75	1.25	0.20	0.75	0.75	0.20
15	1.25	-1.25	0.20	1.25	-0.75	0.20
16	0.75	0.75	0.20	0.75	1.25	0.20
17	1.25	1.25	0.20	1.25	0.75	0.20
18	1.25	0.75	0.20	1.25	1.25	0.20

Note. A, B, and C correspond to the discrimination, difficulty, and psuedo-guessing parameters in the 3PL IRT model for examinees in either the reference (R) or the focal (F) group.

Table 3

Item Parameters for the Large or C-Level DIF Items

Item	$A_R$	$B_R$	$C_R$	$A_F$	$B_F$	$C_F$
1	1.00	0.50	0.20	1.00	-0.50	0.20
2	1.00	-0.50	0.20	1.00	0.50	0.20
3	1.00	1.50	0.20	1.00	0.50	0.20
4	1.00	0.50	0.20	1.00	1.50	0.20
5	1.00	-0.50	0.20	1.00	-1.50	0.20
6	1.00	-1.50	0.20	1.00	-0.50	0.20
7	0.75	0.50	0.20	0.75	-0.50	0.20
8	0.75	-0.50	0.20	0.75	0.50	0.20
9	0.75	-0.50	0.20	0.75	-1.50	0.20
10	1.25	0.50	0.20	1.25	-0.50	0.20
11	0.75	-1.50	0.20	0.75	-0.50	0.20
12	1.25	-0.50	0.20	1.25	0.50	0.20

Note. A, B, and C correspond to the discrimination, difficulty, and psuedo-guessing parameters in the 3PL IRT model for examinees in either the reference (R) or the focal (F) group.

Table 4

Type I Error Rates and Power with 20% DIF (75% Moderate DIF; 25% Large DIF) in the Conditioning Variable as a Function of DIF Method

Ability	Sample Size	Source	Items X Rep	Method						
				<u>MH</u>		<u>SIBTEST</u>		<u>LR</u>		
				Count	Prop.	Count	Prop.	Count	Prop.	
Equal	N <sub>R</sub> =N <sub>F</sub> =250	Type I Error	3200	119	.0372	151	.0472	174	.0544	
		Power	800	599	.7488	621	.7763	583	.7288	
	N <sub>R</sub> =N <sub>F</sub> =500	Type I Error	3200	46	.0144	102	.0319	63	.0197	
		Power	800	675	.8438	755	.9438	662	.8275	
	N <sub>R</sub> =N <sub>F</sub> =1000	Type I Error	3200	5	.0016	6	.0019	8	.0025	
		Power	800	710	.8875	779	.9738	722	.9025	
	Unequal	N <sub>R</sub> =N <sub>F</sub> =250	Type I Error	3200	102	.0319	154	.0481	159	.0497
			Power	800	566	.7075	589	.7363	559	.6988
N <sub>R</sub> =N <sub>F</sub> =500		Type I Error	3200	35	.0109	102	.0319	71	.0222	
		Power	800	650	.8125	728	.9100	651	.8138	
N <sub>R</sub> =N <sub>F</sub> =1000		Type I Error	3200	3	.0009	9	.0028	31	.0097	
		Power	800	653	.8163	758	.9475	724	.9050	

Table 5

Type I Error Rates and Power with 40% DIF (75% Moderate DIF; 25% Large DIF) in the Conditioning Variable as a Function of DIF Method

Ability	Sample Size	Source	Items X Rep	Method					
				<u>MH</u>		<u>SIBTEST</u>		<u>LR</u>	
				Count	Prop.	Count	Prop.	Count	Prop.
Equal	N <sub>R</sub> =N <sub>F</sub> =250	Type I Error	2400	73	.0304	107	.0446	89	.0371
		Power	1600	1186	.7413	1233	.7706	1146	.7163
	N <sub>R</sub> =N <sub>F</sub> =500	Type I Error	2400	22	.0092	66	.0275	25	.0104
		Power	1600	1388	.8675	1511	.9444	1342	.8388
	N <sub>R</sub> =N <sub>F</sub> =1000	Type I Error	2400	0	.0000	2	.0008	4	.0017
		Power	1600	1415	.8844	1558	.9738	1480	.9250
Unequal	N <sub>R</sub> =N <sub>F</sub> =250	Type I Error	2400	77	.0321	115	.0479	102	.0425
		Power	1600	1165	.7281	1203	.7519	1142	.7138
	N <sub>R</sub> =N <sub>F</sub> =500	Type I Error	2400	19	.0079	94	.0392	35	.0146
		Power	1600	1318	.8238	1484	.9275	1310	.8188
	N <sub>R</sub> =N <sub>F</sub> =1000	Type I Error	2400	2	.0008	13	.0054	8	.0033
		Power	1600	1376	.8600	1546	.9663	1454	.9088

Table 6

Type I Error Rates and Power with 60% DIF (75% Moderate DIF; 25% Large DIF) in the Conditioning Variable as a Function of DIF Method

Ability	Sample Size	Source	Items X Rep	Method					
				<u>MH</u>		<u>SIBTEST</u>		<u>LR</u>	
				Count	Prop.	Count	Prop.	Count	Prop.
Equal	N <sub>R</sub> =N <sub>F</sub> =250	Type I Error	1600	51	.0319	66	.0413	57	.0356
		Power	2400	1800	.7500	1860	.7750	1725	.7188
	N <sub>R</sub> =N <sub>F</sub> =500	Type I Error	1600	22	.0138	57	.0356	20	.0125
		Power	2400	1962	.8175	2221	.9254	1949	.8121
	N <sub>R</sub> =N <sub>F</sub> =1000	Type I Error	1600	1	.0006	4	.0025	1	.0006
		Power	2400	2006	.8358	2330	.9708	2163	.9013
Unequal	N <sub>R</sub> =N <sub>F</sub> =250	Type I Error	1600	70	.0438	96	.0600	70	.0438
		Power	2400	1681	.7004	1731	.7213	1658	.6908
	N <sub>R</sub> =N <sub>F</sub> =500	Type I Error	1600	9	.0056	49	.0306	27	.0169
		Power	2400	1866	.7775	2161	.9004	1922	.8008
	N <sub>R</sub> =N <sub>F</sub> =1000	Type I Error	1600	0	.0000	4	.0025	0	.0000
		Power	2400	1875	.7813	2243	.9346	2094	.8725

Table 7

Type I Error Rates and Power with 20% DIF (50% Moderate DIF; 50% Large DIF) in the Conditioning Variable as a Function of DIF Method

Ability	Sample Size	Source	Items X Rep	Method						
				<u>MH</u>		<u>SIBTEST</u>		<u>LR</u>		
				Count	Prop.	Count	Prop.	Count	Prop.	
Equal	N <sub>R</sub> =N <sub>F</sub> =250	Type I Error	3200	110	.0344	153	.0478	143	.0447	
		Power	800	680	.8500	692	.8650	660	.8250	
	N <sub>R</sub> =N <sub>F</sub> =500	Type I Error	3200	38	.0119	78	.0244	60	.0188	
		Power	800	688	.8600	767	.9588	693	.8663	
	N <sub>R</sub> =N <sub>F</sub> =1000	Type I Error	3200	12	.0038	5	.0016	18	.0056	
		Power	800	705	.8813	787	.9838	738	.9225	
	Unequal	N <sub>R</sub> =N <sub>F</sub> =250	Type I Error	3200	115	.0359	158	.0494	147	.0459
			Power	800	598	.7475	602	.7525	605	.7563
N <sub>R</sub> =N <sub>F</sub> =500		Type I Error	3200	43	.0134	117	.0366	73	.0228	
		Power	800	647	.8088	748	.9350	674	.8425	
N <sub>R</sub> =N <sub>F</sub> =1000		Type I Error	3200	7	.0022	17	.0053	20	.0063	
		Power	800	668	.8350	766	.9575	728	.9100	

Table 8

Type I Error Rates and Power with 40% DIF (50% Moderate DIF; 50% Large DIF) in the Conditioning Variable as a Function of DIF Method

Ability	Sample Size	Source	Items X Rep	Method						
				<u>MH</u>		<u>SIBTEST</u>		<u>LR</u>		
				Count	Prop.	Count	Prop.	Count	Prop.	
Equal	N <sub>R</sub> =N <sub>F</sub> =250	Type I Error	2400	90	.0375	116	.0483	94	.0392	
		Power	1600	1346	.8413	1372	.8575	1306	.8163	
	N <sub>R</sub> =N <sub>F</sub> =500	Type I Error	2400	26	.0108	61	.0254	29	.0121	
		Power	1600	1413	.8831	1531	.9569	1399	.8744	
	N <sub>R</sub> =N <sub>F</sub> =1000	Type I Error	2400	0	.0000	4	.0017	2	.0008	
		Power	1600	1458	.9113	1576	.9850	1508	.9425	
	Unequal	N <sub>R</sub> =N <sub>F</sub> =250	Type I Error	2400	98	.0408	137	.0571	124	.0517
			Power	1600	1260	.7875	1305	.8156	1251	.7819
N <sub>R</sub> =N <sub>F</sub> =500		Type I Error	2400	26	.0108	84	.0350	34	.0142	
		Power	1600	1343	.8394	1503	.9394	1376	.8600	
N <sub>R</sub> =N <sub>F</sub> =1000		Type I Error	2400	1	.0004	14	.0058	11	.0046	
		Power	1600	1386	.8663	1558	.9738	1482	.9263	

Table 9

Type I Error Rates and Power with 60% DIF (50% Moderate DIF; 50% Large DIF) in the Conditioning Variable as a Function of DIF Method

Ability	Sample Size	Source	Items X Rep	Method						
				<u>MH</u>		<u>SIBTEST</u>		<u>LR</u>		
				Count	Prop.	Count	Prop.	Count	Prop.	
Equal	N <sub>R</sub> =N <sub>F</sub> =250	Type I Error	1600	63	.0394	79	.0494	63	.0394	
		Power	2400	1984	.8267	2036	.8483	1945	.8104	
	N <sub>R</sub> =N <sub>F</sub> =500	Type I Error	1600	11	.0069	48	.0300	14	.0088	
		Power	2400	2187	.9113	2317	.9654	2162	.9008	
	N <sub>R</sub> =N <sub>F</sub> =1000	Type I Error	1600	0	.0000	0	.0000	0	.0000	
		Power	2400	2236	.9317	2361	.9838	2288	.9533	
	Unequal	N <sub>R</sub> =N <sub>F</sub> =250	Type I Error	1600	56	.0350	65	.0406	63	.0394
			Power	2400	1935	.8063	2005	.8354	1903	.7929
N <sub>R</sub> =N <sub>F</sub> =500		Type I Error	1600	18	.0113	71	.0444	20	.0125	
		Power	2400	2137	.8904	2282	.9508	2145	.8938	
N <sub>R</sub> =N <sub>F</sub> =1000		Type I Error	1600	2	.0013	10	.0063	3	.0019	
		Power	2400	2151	.8963	2340	.9750	2233	.9304	

Figure Caption

Figure 1. The 2 X 2 contingency table formed for each level of the matching criterion, total test score, as used with Mantel-Haenszel.

		<u>Score on Studied Item</u>		
		1	0	Total
<u>Group</u>	Reference Group	$A_j$	$B_j$	$N_{r_j}$
	Focal Group	$C_j$	$D_j$	$N_{f_j}$
	Total	$T_{1j}$	$T_{0j}$	$T_j$

Note. The subscript  $j$  refers to the score level,  $i$  to the test item,  $r$  to the reference group, and  $f$  to the focal group.