

**Using Real Data to Compare DIF Detection and Effect Size Measures among
Mantel-Haenszel, SIBTEST, and Logistic Regression Procedures**

Yinggan Zheng

Mark J. Gierl

Ying Cui

Centre for Research in Applied Measurement and Evaluation

University of Alberta

6-110 Education North

Edmonton, AB, Canada T6G2G5

Abstract

To date, many studies have been conducted to compare the performance of different DIF procedures using simulated data sets. However, some results from these simulation studies are inconsistent with one other (e.g., Hidalgo & López-Pina, 2004; Jodoin & Gierl, 2001). This study used real data to systematically investigate the consistencies of DIF detection and effect size among three widely-used DIF procedures: Mantel-Haenszel (MH), Simultaneous Item Bias Test (SIBTEST), and logistic regression (LR). Several indicators, including correlations among DIF procedures effect size measures, matching percentages, and relative matching percentages, were used to evaluate the consistencies among the DIF procedures. The results showed high correlations among DIF effect size measures, moderate to high matching percentages among DIF classifications, and a broad range of relative matching percentages among DIF procedures.

Using Real Data to Compare DIF Detection and Effect Size Measures among Mantel-Haenszel, SIBTEST, and Logistic Regression Procedures

Differential item functioning (DIF) is of great interest to researchers and educators given that DIF poses a potential threat to test fairness. A variety of DIF detection procedures and effect size measures have been proposed to quantify the magnitude of DIF, such as the IRT methods (Lord, 1980; Thissen, Steinberg, & Wainer, 1993), the Mantel-Haenszel statistic (MH; Holland & Thayer, 1988), the standardization procedure (Dorans & Kullick, 1986), the Simultaneous Item Bias Test (SIBTEST; Shealy & Stout, 1993a), and Logistic Regression (LR; Swaminathan & Rogers, 1990). Often, multiple procedures are used simultaneously to help detect items showing DIF (Hambleton & Jones, 1994). As a result, high consistencies or matching percentages among DIF procedures are of consequence in real testing situation. To date, many studies have been conducted to compare the performance of different DIF procedures using simulated data sets (e.g., Fidalgo, Ferreres, & Muniz, 2004; Gierl, Jodoin, & Ackerman, 2000; Hidalgo & López-Pina, 2004; Jodoin & Gierl, 2001; Narayanan & Swaminathan, 1994; Roussos & Stout, 1996). However, some results from these simulation studies are inconsistent with one other. For instance, Jodoin and Gierl (2001) reported that 68.2% of DIF items were identified as containing at least moderate DIF using logistic regression procedure in their simulation study. In contrast, Hidalgo and López-Pina (2004) found that, by using the same procedure and same classification guideline, only 15.3% of DIF items were classified as having moderate DIF. Therefore, the purpose of this study is to use real data to systematically investigate the consistencies of DIF detection and effect size measures among three widely-used DIF procedures: the MH procedure, the

SIBTEST procedure, and the LR procedure. This paper is divided into four sections. First, a brief overview of the three DIF procedures is provided. Second, the methods used in this study for evaluating the consistencies among the three DIF procedures are described. Third, the results are presented. Forth, the implications and future directions are discussed.

Overview of the three DIF procedures

Mantel-Haenszel

Mantel-Haenszel (MH) is one of the widely-used approaches for identifying DIF based on analysis of contingency tables (Clauser & Mazor, 1998; Holland & Thayer, 1988). In MH procedure, a chi-square test with one degree of freedom is yielded to test the null hypothesis that there is no relation between group membership and test performance on one item after controlling for ability. MH is computed by matching examinees in each group on their total test scores and then forming a 2-by-2-by- K contingency table for each item, where K is the total number of score levels on the matching variable, namely, the 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.

The MH chi-square test is calculated as follows:

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

where,

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

and,

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

The MH procedure also estimates the constant odds ratio that yields a measure of effect size for evaluating the magnitude of DIF. The odds ratio is calculated as follows:

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

The α_{MH} is the ratio of the odds that a reference group examinee will get the item correct compared to the odds for a matched focal group examinee. The α_{MH} is often transformed to the Δ_{MH} to enhance the interpretability of the result using the formula,

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

Based on this transformation, Zwick and Ercikan (1989) proposed the following interpretation guidelines to evaluate the DIF effect size:

- Negligible or A-level DIF: $|\Delta_{MH}| < 1$, or MH test is not statistically significant,
- Moderate or B-level DIF: $1 \leq |\Delta_{MH}| < 1.5$ and MH test is statistically significant,
- Large or C-level DIF: $|\Delta_{MH}| \geq 1.5$ and MH test is statistically significant.

Simultaneous Item Bias Test (SIBTEST)

The Simultaneous Item Bias Test (SIBTEST) is an alternative statistical method for detecting DIF proposed by Shealy and Stout (1993). The null hypothesis tested by SIBTEST is:

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

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 respective value for examinees in the Focal group with true score T . The test statistic for measuring the null hypothesis in SIBTEST is:

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

where,

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

and,

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

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 using a regression correction procedure outlined in Shealy & Stout (1993); $\hat{\sigma}^2(Y|k, g)$ is the sample variance for examinees on the studied item 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.

Based on the highly correlated relationship between Δ_{MH} and $\hat{\beta}$ (an estimate of

the amount of DIF in SIBTEST), Roussos and Stout (1996, p. 220) proposed the following interpretation guidelines to evaluate the DIF effect size:

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

Logistic Regression DIF

Swaminathan and Rogers (1990) applied the logistic regression (LR) procedure, a model-based approach, to identify DIF. The equation in LR model for DIF detection is expressed as

$$P(u = 1 | \theta, g) = \frac{e^{f(\theta, g)}}{1 + e^{f(\theta, g)}},$$

where $P(u = 1 | \theta, g)$ is the conditional probability of obtaining a correct answer given the vector of independent variables (i.e., θ, g). $f(\theta, g)$ is the function that defines the linear combination of the predictor variables, including the observed ability (θ), the group membership (g), and the interaction between the observed ability and the group membership (θg). The $f(\theta, g)$ can be expressed dependent on the steps in the LR procedure. In step 1, $f(\theta, g)$ equal to $\tau_0 + \tau_1\theta$ (model 1), where the coefficients τ_0, τ_1 represent the intercept and weights for the ability. This serves as the baseline model. In step 2, 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., $f(\theta, g) = \tau_0 + \tau_1\theta + \tau_2g$) subtracted from Model 1. In step 3, 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 (θg) against model 2. That is, Model 3 (i.e. $f(\theta, g) = \tau_0 + \tau_1\theta + \tau_2g + \tau_3\theta g$) subtracted from Model 2.

Jodoin and Gierl (2001) recently evaluated the use of an effect size measure for uniform DIF detection, called $R^2\Delta - U$, with logistic regression in an attempt to reduce the inflated Type I errors often associated with this approach (Narayanan & Swaminathan, 1996; Swaminathan & Rogers, 1990). $R^2\Delta - U$ is given as:

$$R^2\Delta - U = R_2^2 - R_1^2,$$

where R_2^2 and R_1^2 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., $\sum_1^j \beta_j r_j$ for j explanatory variables) of the model 2 and model

1. They 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. The guidelines are:

- Negligible or A-level DIF: $\Delta R^2 < 0.035$
- Moderate or B-level DIF: Null hypothesis is rejected and $0.035 \leq \Delta R^2 < 0.07$
- Large or C-level DIF: Null hypothesis is rejected and $\Delta R^2 \geq 0.07$

Method

Data

In this study, data from two large-scale assessments were used for language and gender DIF analyses. Data used for language DIF analyses consist of three subject areas, computer studies, physics, and history, each with two language versions, English and Chinese. Within each reference (English) and focal (Chinese) group, a sample of 2000

examinees was randomly selected. Data, across three grades (grade 3, 6, and 9) and three subject areas (science, math, and social studies), were used for gender DIF analyses. A sample of 2000 examinees was randomly selected for each reference (male) and focal (female) group.

Data analyses

In order to correctly identify negligible, moderate, and large DIF items, classification guidelines must be used for each DIF effect size measure associated with the DIF procedures. For MH, the guideline proposed by Zwick and Ercikan (1989) was used. The guideline developed by Roussos and Stout (1996b) was employed for SIBTEST. The guideline introduced by Jodoin and Gierl (2001) was used for LR. In this study, only uniform DIF (i.e., the interaction between groups and ability levels is not statistically significant) was detected by LR procedure.

Three indicators were used to examine the extent to which different DIF procedures are consistent with one another. These indicators include correlations among the effect size measures across the three procedures, matching percentages, and relative matching percentages. The correlation between the effect size measures of two procedures can help answer the questions: Is there a linear relationship between these effect size measures? If yes, what are the magnitude and direction of this relationship? A high positive correlation indicates that these two procedures provide consistent results in identifying the magnitude and direction of DIF. The matching percentage between two procedures is used to answer the question: Of the items that are identified by at least one of the two compared procedures, how many items are identified by both procedures? A high matching percentage suggests that the two procedures are consistent in terms of

identifying DIF items. Relative matching percentage can answer the question: Of all the items identified by the studied procedure, how many items are also identified by another DIF procedure? A high relative matching percentage indicates that the studied procedure tends to have a low type I error rate.

Results

Translation DIF

Correlations among effect size measures

Table 1 presents the correlations among the three effect size measures across the three subject areas, including computer studies, physics, and history. For computer studies, a high correlation, 0.90, was found between the SIBTEST effect size measure (\hat{B}) and the MH effect size measure ($\Delta - MH$). \hat{B} was also highly correlated with the LR uniform DIF effect size measure ($R^2\Delta - U$), 0.91. A moderate correlation, $r = 0.79$, was found between $\Delta - MH$ and $R^2\Delta - U$. High correlations were found among the three effect size measures for physics, 0.96 between $\Delta - MH$ and \hat{B} , 0.91 between $\Delta - MH$ and $R^2\Delta - U$, and 0.86 between \hat{B} and $R^2\Delta - U$. Similarly, the subject area of history produced high correlations among the three effect size measures. The MH effect size measure showed strong correlations with \hat{B} ($r = 0.97$) as well as with $R^2\Delta - U$ ($r = 0.93$). There was also a strong correlation between \hat{B} and $R^2\Delta - U$ ($r = 0.92$).

Matching percentages among three DIF procedures

Translation DIF classifications were conducted based on effect size measures using three procedures for each of three subject areas, computer studies, physics, and history. In this section, items with a B- or C-level rating are considered as DIF items

whereas those with an A-level rating are not. The number of identified DIF items and the number of matching DIF items across the three procedures are summarized in Table 2. Classification percentages and matching percentages are also presented in Table 2. For classification percentage, the denominator is the number of total items, while the numerator is the number of items identified as displaying DIF. For matching percentage, the denominator is the number of items identified by at least one of two procedures under comparison, while the numerator is the number of items identified as showing DIF by both of the procedures under comparison. To compare the consistency between MH and LR, for example, five physics items were identified by MH (items 3, 4, 27, 42, and 46), while six items by LR (items 3, 4, 27, 42, 45, and 46). There were six items in total that were identified as showing DIF either by MH or LR or both. Therefore, the denominator of the matching percentage between MH and LR was six. Of the six items, five items were identified as showing DIF by both MH and LR (items 3, 4, 27, 42, and 46). Then the numerator of the matching percentage between MH and LR was five. As a result, the matching percentage was $5/6 = 83.30\%$.

For computer studies, as shown in Table 2, the SIBTEST procedure identified the largest number of DIF items (30), 62.5% of the total 48 items, followed by LR (27 items), 56.25% of the total items. MH identified the fewest number of items (24). The matching percentage between MH and SIBTEST is 68.75%, obtained by using the number of items identified as showing DIF by both MH and SIBTEST (22) divided by the total number of items identified at least by one of these two procedures (32). The matching percentages between MH and LR-U and between SIBTEST and LR-U were 75.86% and 90%, respectively.

For physics, SIBTEST identified the largest number of DIF items (10), followed by LR (6), and MH (5). The matching percentages between MH and SIBTEST, between MH and LR-U, and between SIBTEST and LR-U were, 50.00%, 83.30%, and 60.00%, respectively.

For history, SIBTEST procedure identified the largest number of DIF items, 14, 36.84% of the total items. Procedure LR identified twelve (31.58%) DIF items, and MH identified nine DIF items (23.68%), respectively. Nine items (64.29%) were matched between MH and SIBTEST, nine items (75.00%) between MH and LR-U, and eleven items (73.33%) between SIBTEST and LR-U.

Relative matching percentages among the three procedures

In order to evaluate the consistency among the three procedures in detecting DIF item, relative matching percentage (defined as proportion of matching DIF items based on studied procedure) was used. In Table 3, each row represents the performance of the corresponding studied procedure. For instance, the first row of Table 3 shows the performance of MH procedure. Of the total 24 DIF items identified by MH, 22 (91.67%) items were also identified by SIBTEST and LR-U, respectively. According to the second row of Table 3, 22 of 30 DIF items (73.33%) identified by SIBTEST were also identified by MH, and 27 (90.00%) items by LR-U as displaying uniform DIF. Similarly, the last row shows that 22 of 27 items (81.48%) identified by LR-U were also identified by MH, and 27 items (100.00%) by SIBTEST. For physics and history, the results can be interpreted in the same manner.

Gender DIF

Correlations among the effect size measures

For the data of achievement tests, gender DIF analyses were conducted. In the area of science, the correlations among the effect sizes measures across the three procedures ranged from 0.66 to 0.96 (Mean=0.85, SD=0.11) across grades 6, 9, and 12. For mathematics, the correlations ranged from 0.84 to 0.96 (Mean=0.91, SD=0.04). For social study, the correlations ranged from 0.87 to 0.93 (Mean=0.90, SD=0.02).

Matching percentages among the three procedures

Table 4 shows the number and percentage of items identified as showing DIF by each procedure. In addition, Table 4 also presents the number and percentage of DIF items that were identified consistently by each pair of procedures under consideration. For example, of the total 56 grade-12 science items, 2 (3.57%), 5 (8.93%), and 5 (8.93%) items were identified as displaying DIF by MH, SIBTEST, and LR-U, respectively. Two items were consistently identified as showing DIF by MH and SIBTEST. The associated matching percentage was 40%. Likewise, the matching percentages between MH and LR-U and between SIBTEST and LR-U were 40% and 66.67%, respectively.

Relative matching percentages among the three procedures

Table 5 displays the number of matching items identified as showing DIF by procedures across all the subjects, along with the relative matching percentages associated with pair of procedures. For the MH procedure, highest relative matching percentages appear (Mean=94.82%, SD=14.82%) partly because it identified the fewest number of items as showing DIF. Lowest relative matching percentages were found for SIBTEST (Mean=51.60%, SD=21.09%). LR-U had the middle matching percentages in this study (Mean=63.61%, SD=22.93%).

Implications and Future Research

This study used real data to systematically compare DIF detection and effect size measures among three frequently used DIF procedures. Several indicators, including correlations among DIF procedures effect size measures, matching percentages, and relative matching percentages, were used to evaluate the consistencies among the DIF procedures.

The results of this study have several implications for practice. First, high positive correlations were found among different DIF effect size measures, suggesting that different procedures provided consistent estimates on the magnitude and direction of DIF. Second, moderate to high matching percentages were found among procedures, indicating that the classification consistencies among procedures are relatively strong but not perfect. This finding supports many researchers' recommendation of the use of multiple DIF procedures in real testing situation to reduce the uncertainty associated with the analysis of empirical data. Third, the highest matching percentages were found using guidelines for LR developed by Jodoin and Gierl (2001) and for SIBTEST developed by Roussos and Stout (1996b). This result suggests that the guideline developed by Jodoin and Gierl (2001) provides a reliable classification of DIF items. In future research, simulation studies should be conducted for two purposes. The first purpose is to evaluate the consistencies among DIF procedures using the three indicators under different manipulated conditions. The second purpose is to systematically evaluate the use of multiple procedures by examining whether items that are consistently identified by different combination of procedures truly contain DIF.

Reference

- Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased items*. Thousand Oaks, 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.
- Dorans, N. J., & Kullick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23*, 355-368.
- Fidalgo, A. M., Ferreres, D., Muniz, J., (2004). Liberal and conservative differential item functioning detection using Mantel-Haenszel and SIBTEST: implications for Type I and Type II error rates. *The Journal of Experimental Education, 2004, 7J(I)*, 23-39
- Gierl, M. J., Jodoin, M. G., & Ackerman, T. A. (2000, April). *Performance of Mantel-Haenszel, Simultaneous Item Bias Test, and Logistic Regression When the Proportion of DIF Items is Large*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.
- Gierl, M. J., Rogers, W. T., & Klinger, D. (1999). Using statistical and substantive reviews to identify and interpret translation DIF. *Alberta Journal of Educational Research, 45*, 353-376.
- Hambleton, R. K., & Jones, R. W. (1994). Comparison of empirical and judgmental procedures for detecting differential item functioning. *Educational Research Quarterly, 18*, 21-36.

- Hidalgo, M. D. & Lopez-Pina, J. A. (2004). Differential item functioning detection and effect size: a comparison between logistic regression and Mantel-Haenszel procedures. *Educational and Psychological Measurement, Vol. 64 No. 6*, 903-915
- 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. (2001). Evaluating Type I error and power rates using an effect size measure with logistic regression procedure for DIF detection. *Applied Measurement in Education, 14*, 329-349.
- Li, H., & Stout, W. (1996). A new procedure for detection of crossing DIF. *Psychometrika, 61*, 647-677.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.
- Mazor, K. M., Kanjee, A., Clauser, B. E. (1995) Using logistic regression and the Mantel-Haenszel with multiple ability estimates to detect differential item functioning. *Journal of Educational Measurement, Vol 32(2) Sum 1995*, 131-144.
- 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.

- 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. T., & Stout, W. F. (1993a). An item response theory model for test bias and differential test functioning. In P. Holland & H. Wainer (Eds.) *Differential Item Functioning* (pp. 197-239). Hillsdale, NJ.
- Shealy, R. T., & Stout, W. F. (1993b). 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, 54*, 159-194.
- Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*, 361-370.
- Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67-114). Hillsdale, NJ: Erlbaum.
- Zwick, R., & Ercikan, K. (1989). Analysis of differential item functioning in the NAEP history assessment. *Journal of Educational Measurement, 26*, 55-66.
- 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.

Table 1.
Correlations among Effect Size Measures across Subject Areas for Translation DIF

	<u>Computer Study</u>			<u>Physics</u>			<u>History</u>		
	Δ -MH	\hat{B}	$R^2\Delta-U$	Δ -MH	\hat{B}	$R^2\Delta-U$	Δ -MH	\hat{B}	$R^2\Delta-U$
Δ -MH	—			—			—		
\hat{B}	0.90	—		0.96	—		0.97	—	
$R^2\Delta-U$	0.79	0.91	—	0.91	0.86	—	0.93	0.91	—

Table 2.

Classification and Consistency among Procedures across Subjects for Translation DIF

	Classification			Consistency		
	MH	SIBTEST	LR-U	MH& SIBTEST	MH& LR-U	SIBTEST &LR-U
Computer Studies	24(50.00%)	30(62.50%)	27(56.25%)	22(68.75%)	22(75.86%)	27(90.00%)
Physics	5(11.11%)	10(22.22%)	6(13.30%)	5(50.00%)	5(83.30%)	6(60.00%)
History	9(23.68%)	14(36.84%)	12(31.58%)	9(64.29%)	9(75.00%)	11(73.33%)

Table 3.

Relative Matching Percentages among Procedures across Subjects for Translation DIF

	<u>Computer Study</u>			<u>Physics</u>			<u>History</u>		
	MH	SIBTEST	LR-U	MH	SIBTEST	LR-U	MH	SIBTEST	LR-U
MH	—	22(91.67%)	22(91.67%)	—	5(100.00%)	5(100.00%)	—	9(100.00%)	9(100.00%)
SIBTEST	22(73.33%)	—	27(90.00%)	5(50.00%)	—	6(60.00%)	9(64.29%)	—	11(78.57%)
LR-U	22(81.48%)	27(100.00%)	—	5(83.33%)	6(100%)	—	9(75%)	11(91.67%)	—

Table 4

Classification and Consistency among Procedures across Subjects and Grades for Gender DIF

		Classification			Consistency		
		MH	SIBTEST	LR-U	MH& SIBTEST	MH& LR-U	SIBTEST &LR-U
Grade 12	Science	2(3.57%)	5(8.93%)	5(8.93%)	2(40.00%)	2(40.00%)	4(66.67%)
	Math	2(5.00%)	4(10.00%)	3(7.50%)	2(50.00%)	2(66.67%)	2(40.00%)
	Social Study	9(12.86%)	19(27.14%)	17(24.29%)	8(40.00%)	7(36.84%)	16(80.00%)
Grade 9	Science	6(10.91%)	12(21.82%)	10(18.18%)	6(50.00%)	6(60.00%)	9(69.23%)
	Math	2(5.00%)	5(12.50%)	5(12.50%)	2(40.00%)	2(40.00%)	4(66.67%)
	Social Study	1(1.82%)	7(12.73%)	6(10.91%)	1(14.29%)	1(16.67%)	4(44.44%)
Grade 6	Science	3(6.00%)	9(18.00%)	8(16.00%)	3(33.33%)	3(37.50%)	7(70.00%)
	Math	5(10.00%)	5(10.00%)	7(14.00%)	2(25.00%)	5(71.43%)	4(50.00%)
	Social Study	1(2.00%)	5(10.00%)	2(4.00%)	1(20.00%)	1(50.00%)	2(40.00%)

Table 5.

Relative Matching Percentages among Procedures across Subjects & Grades for Gender DIF

		<u>Science</u>			<u>Math</u>			<u>Social Study</u>		
		MH	SIBTEST	LR-U	MH	SIBTEST	LR-U	MH	SIBTEST	LR-U
Grade 12	MH	—	2(100.00%)	2(100.00%)	—	2(100.00%)	2(100.00%)	—	8(88.89%)	7(77.78%)
	SIBTEST	2(40.00%)	—	4(80.00%)	2(50.00%)	—	2(50.00%)	8(42.11%)	—	16(84.21%)
	LR-U	2(40.00%)	4(80.00%)	—	2(66.67%)	2(66.67%)	—	7(41.18%)	16(94.12%)	—
Grade 9	MH	—	6(100.00%)	6(100.00%)	—	2(100.00%)	2(100.00%)	—	1(100.00%)	1(100.00%)
	SIBTEST	6(50.00%)	—	6(50.00%)	2(40.00%)	—	4(80.00%)	1(14.29%)	—	4(57.14%)
	LR-U	6(60.00%)	9(90.00%)	—	2(40.00%)	4(80.00%)	—	1(16.67%)	4(66.67%)	—
Grade 6	MH	—	3(100.00%)	3(100.00%)	—	2(40.00%)	5(100.00%)	—	1(100.00%)	1(100.00%)
	SIBTEST	3(33.33%)	—	7(77.78%)	2(40.00%)	—	4(80.00%)	1(20.00%)	—	2(40.00%)
	LR-U	3(37.50%)	7(87.50%)	—	5(71.43%)	4(57.14%)	—	1(50.00%)	2(100.00%)	—

		<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.

Figure 1. The 2 X 2 Contingency Table Formed for Each Level of the Matching Criterion, Total Test Score, as Used with the Mantel-Haenszel Procedure.