The Effects of Matching Type and Sample Size on the Mantel-Haenszel Technique for Detecting Items with DIF

Zilberberg, A., Phan, H., Socha, A., Kong, J., & Keng, L

Federal mandates for educational accountability (No Child Left Behind Act, 2002) led to the proliferation of large-scale assessment in K-12 settings. Expectedly, state-mandated assessments undergo a rigorous test development process, during which the best efforts are made by content experts and psychometricians to ensure fairness and equity in testing. As part of this process, statistical analyses of examinees' scores are conducted in order to detect items that *function differentially* (i.e., items with DIF) for members of minority groups that are comparable in ability with the members of the majority group (Dorans & Holland, 1993). DIF occurs when groups of test-takers of the same ability have different probabilities of answering the test item correctly (Zeiky, 1993). In such studies, relative performance of the *focal group* (e.g., ethnic minority, disability group, females, etc.) is compared to that of the *reference* group (e.g., Caucasians, typically-functioning, males, etc.). In practice, when test items are identified as having DIF they are often presented at data review meetings at an early stage in the test development process so that content experts can review these items for additional evidence of DIF and make informed decisions about whether or not to include such items in the item bank.

One non-parametric method for statistical detection of DIF is the Mantel-Haenszel (MH) procedure. First proposed as a DIF detection method for dichotomously-scored items by Holland and Thayer in 1988, the MH procedure is based on the contingency table, with counts of correct (1) and incorrect (0) responses broken up by the group indicator (focal and reference groups) and the matching criterion (*j* categories). The MH procedure is based on comparing *matched* groups, so that item functioning can be evaluated conditional on ability (e.g., total score). The total score is most commonly used as the matching criterion. The $\hat{\alpha}_{MH}$ odds-ratio estimator is obtained using formula (1), where *a* and *c* are the number of students who answer the item correctly in reference and focal groups, respectively, and *d* and *b* are the number of students who answer the item incorrectly in the reference and focal groups, respectively.

$$\hat{\alpha}_{MH} = \frac{\sum_{j=1}^{k}\left(\frac{a_j d_j}{n_j}\right)}{\sum_{j=1}^{k}\left(\frac{b_j c_j}{n_j}\right)} \tag{1}$$

$$\hat{\beta} = \ln\left(\hat{\alpha}_{MH}\right) \tag{2}$$

$$\Delta_{MH} = -2.35\hat{\beta} \tag{3}$$

The signed index is $\hat{\beta}$ obtained by taking the natural log of common-odds ratio, as demonstrated in formula (2). The Δ_{MH} is obtained by multiplying the signed index by -2.35, as demonstrated in formula (3). This index is used to supplement the sample-sensitive χ^2 test (used with the $\hat{\alpha}_{MH}$ odds-ratio). Positive values of the Δ_{MH} indicate that the item favors the focal group, whereas negative values indicate that the item disadvantages the focal group. Both MH χ^2 and the absolute value of Δ_{MH} are recommended for detecting DIF so that the amount of DIF can be classified as negligible, moderate, or high (Zeiky, 1993). Negligible, moderate, or high DIF items are also widely referred to as the ETS DIF classified A, B, and C items, respectively (see Appendix; Zieky, 1993; Zwick & Ercikan, 1989).

One issue related to DIF detection using MH is sample size. Schmitt, Holland, and Dorans (1993) suggest that, whenever feasible, the largest possible sample sizes of both focal and reference groups should be used in the DIF analyses. However, given that minority groups are the focus of DIF analyses, small sample sizes and thus under-powered studies are often inevitable. Some argue for using even relatively small sample sizes (*N* = 100) in DIF analyses, "weighing the harm that could be caused by relatively unstable statistics against the harm that could be caused by failure to do any analyses at all" (Zeiky, p. 345). Yet other researchers caution against using multiple small samples due to accumulation of Type I error (e.g., Linn, 1993). Moreover, simulation studies indicate that such small sample sizes lead to inadequate recovery of DIF and underpowered statistics (Schultz, Perlman, Rice, & Wright, 1989; Swaminathan & Rogers, 1990). In addition, some evidence indicates that the MH indices function poorly if a focal group is small and a group separation is large (Camilli & Smith, 1990). However, recent investigation of the MH performance in the situations when the focal and reference groups have asymmetrically unbalanced but large sample sizes reveals adequate results (Paek & Guo, 2011). In sum, variations in groups' sample sizes pose a technical challenge to detecting items with DIF because DIF statistics become less stable as sample sizes decrease (Zeiky, 1993).

Determining matching categories is an important step in the MH procedure (Donoghue & Allen, 1993; Holland & Thayer, 1988; Zwick, 1991). Although a crude approximation of the ability distribution, matching nonetheless provides a useful way to group examinees according to their ability (Scheuneman, 1979). The categories can be determined by discretizing the total score, including the studied item, into a number of score ranges. However, the answer pertaining to the optimal way of determining these categories is unclear. Scheuneman (1979) offers the following criteria for creating ability categories: (1) probability of a correct response in each category should be less than 1; (2) expected frequency in each category should be at least five; and (3) the smallest observed cell frequency should be about the same at each ability level. Various ways of creating ability intervals based on the total score were also examined by other researchers. Donoghue and Allen (1993) examined *thin matching* and *thick matching* and the effects of each on the MH technique. In *thin matching,* "each total test score is a separate ability level within which focal and comparison group students are expected to have equal probability of correctly answering the item under study" (Schulz, Perlman, Rice, & Wright, 1996, p. 67). Donoghue & Allen (1993) demonstrated that *thin matching* yielded poor results when compared to *thick matching*. Various types of *thick matching*, or grouping score ranges into the ability intervals, are available, but research on the interaction of the effects of different matching types, ability levels, and sample sizes on the MH technique is inconclusive (Donoghue & Allen, 1993; Scheuneman, 1979).

The purpose of this simulation study is to investigate the effects of sample size and matching type on how well the MH procedure detects test items that function differentially (DIF) across different groups of students (i.e., gender and ethnicity groups) of similar ability. This simulation study will use real item parameters from a recent large-scale state-mandated high-school level mathematics test to simulate test items for the study. It is hoped that results of the simulation study would be used to inform policy-related decisions around DIF analysis of assessment items such as the minimum sample size requirement for a focal group and the recommended matching type for a DIF analysis to yield reliable results.

Method

The data for this study was simulated and analyzed in SAS 9.2 (SAS Institute, 2008) using PROC IML. Item parameters for the simulated test were based on a real state-mandated high-school mathematics test which includes 62 multiple-choice items and has an overall item mean Rasch value of .487 and variance of .241. Data were generated to follow a 1PL model, with item difficulty (*b*) parameters following a normal distribution with this mean and standard deviation.

Four different matching types were manipulated in the study: quartiles, deciles, two-adjacent (combining every two adjacent scores) and four-adjacent (combining every four adjacent scores). Matching was done on raw scores tallied from the focal and reference groups after DIF was introduced. Table 1 provides the descriptions of the four above-mentioned matching types and Table 2 lists a total of 12 sample size conditions that were simulated, with the focal group comprising from 5% to 50% of the total sample.

Table 1. Matching Type

Matching Type	Description
Quartiles	Total score dived into 4 percentile groups
Deciles	Total score divided into 10 percentile groups
Two-adjacent	Every two adjacent scores are collapsed to make 31 groups
Four-adjacent	Every four adjacent scores are collapsed to make 15 groups

Table 1. Sample Size Conditions

% in Focal Group	# in Focal Group	# in Reference Group
	Sample Size = 500	
5%	25	475
10%	50	450
25%	125	375
50%	250	250
	Sample Size = 1,000	
5%	50	950
10%	100	900
25%	250	750
50%	500	500
	Sample Size = 2,000	
5%	100	1900
10%	200	1800
25%	500	1500
50%	1000	1000

There were two conditions under which sample size and matching types outlined above were manipulated: (1) baseline condition and (2) DIF condition. In the baseline condition, there are 62 items and DIF was not introduced to any of the items. In the DIF condition, only unidirectional DIF was introduced, so that each DIF item favored the reference group (i.e., more difficult for the focal group). Three levels of ETS-classified DIF were introduced to the simulated difficulty parameters of 10 randomly selected items on the test and the 52 remaining items were set to be free of DIF. Of the 10 DIF items, three had negligible or A level DIF, four had moderate or B level DIF, and three had large or C level DIF. Based on the real data, A level DIF was defined as an average difference in Rasch item difficulties of .20 logit between the reference group and focal groups; B level DIF was defined as an average difference in Rasch item difficulties of .45 logit; and C level DIF was defined as an average difference in Rasch item difficulties of .78 logit. Table 3 presents changes in the *b*-parameters and the corresponding ETS-classified DIF level for the 10 items generated to have DIF.

Table 3. Three Levels of DIF Introduced to 10 Items of the Simulated Test

Item	Change in *b* parameter*	ETS DIF Category
3	0.2	A
28	0.2	A
52	0.2	A
21	0.45	B
26	0.45	B
43	0.45	B
50	0.45	B
6	0.78	C
23	0.78	C
37	0.78	C

*****Values added to the 10 items generated item difficulties

Finally, the simulated ability distributions of the reference and focal groups are normally distributed with a mean 0 and standard deviation of 1. For each condition, 100 test response files (i.e., replications) were generated. This number of replications was simulated to obtain more stable estimates and increase variability across samples. Moreover, 100 iterations is a common practice when conducting DIF and comparability studies for multiple-choice tests.

In order to determine if a test item displays DIF, both statistical and practical significance were considered. For statistical significance, confidence intervals around Δ_{MH} were used and for practical significance, the ETS DIF classification rules were used. Items were flagged as showing a significant DIF if classified as “B” category or above. A flowchart illustrating the decision rules for classifying test items into one of the three ETS-classified DIF levels A, B, and C can be found in Appendix F.

Results

Baseline Condition

Appendix A presents descriptive statistics for estimated Δ_{MH} and true positive rate for MH test, which is the proportion of replications classified as not having DIF, by each of the sample size, percentage of sample in the focal group, and matching type conditions for the baseline condition (i.e., not items generated with DIF). Figure 1 displays these results graphically.

As expected, larger total sample size yielded better classification rates across all matching types. Specifically, correct classification rate ranged from 54.29% to 93.05% (average = 71.94%) when the total sample size equaled 500; from 71.71% to 98.90% (average = 88.56%) when total sample size equaled 1,000; and from 87.58% to 99.90% (average = 95.95%) when total sample size equaled 2,000. When the total sample size was smaller, balanced designs, in which focal group approximated the reference group in size, yielded better classification rates. For example, when the total sample size was 500, the average classification rate (across matching types) with the focal group comprising 5% of the total sample was 55.59%; it increased to 70.07% when the focal group comprised 10% of the sample; it increased to 87.78% when the focal group comprised 25% of the sample; and it further increased to 92.09% when the focal group comprised 50% of the sample. A similar pattern was observed in the condition where the total sample size was 1,000. However, the classification rates in this condition were better overall (ranging from 71.71% to 98.90%) and the added benefits of symmetric groups were less evident. When the sample size equaled 2,000, the overall classification rate was even better, across matching types and focal group sizes.

Differences across matching types appear to be minor in the baseline condition. Quartiles yielded a higher average classification rate (across sample sizes) of 84.52%, followed by deciles (83.79%), followed by four-adjacent (83.60%), followed by two-adjacent (83.09%). The differences among matching types were most evident in the smaller sample size conditions. However, when the total sample size reaches 1,000 with the focal group comprising at least 25% of the total sample, the differences among the four matching types diminish and they perform approximately equally well.

It is important to note that Δ_{MH} could not be calculated for one replication across all matching types in the condition with a sample size of 500 and 5 percent of this sample in the focal group. It is surprising that more replications did not converge since examinees whose responses are presented in incomplete categories, such as when the sample size is small and the number of items (and thus score categories) is large, may be lost from calculations.

DIF Condition

Appendix B presents descriptive statistics for estimated Δ_{MH} and true positive rate for MH test (the proportion of replications having a significant value of MH χ^2), by each of the sample size, percentage of sample in the focal group, and matching types for the 52 items generated to be free of DIF. Appendices C, D, and E present descriptive statistics for estimated Δ_{MH} and true positive rate for MH test for the conditions where A-level DIF was introduced to 3 items, B-level DIF was introduced to 4 items, and C-level DIF was introduced to 3 items, respectively. An item was classified as having DIF if it met both statistical and practical significance criteria. Figure 2 displays these results graphically.

Similarly to the baseline condition and expectedly so, larger total sample size yielded better classification rates across all matching types and ETS DIF categories. Specifically, correct classification rate ranged from 11.50% to 100% (average = 67.40%) when the total sample size equaled 500; from 19.50% to 100% (average = 82.28%) when total sample size equaled 1,000; and from 25.50% to 100% (average = 90.56%) when total sample size equaled 2,000. For the items that had no DIF or A-level DIF, the classification rates remained in the acceptable range, ranging from the average of 69.65% (with the total sample size of 500) to 85.44% (with the total sample size of 1,000) to 93.95% (with the total sample size of 2,000). Similarly, classification rates were also adequate for the items that had C-level DIF, ranging from the average of 90.35% (with the total sample size of 500) to 95.71% (with the total sample size of 1,000) to 98.92% (with the total sample size of 2,000). However, classification rates for B-level items were subpar, ranging from the average of 19.30% (with the total sample size of 500) to 28.73% (with the total sample size of 1,000) to 37.73% (with the total sample size of 2,000).

Matching types, ignoring the effects of sample size and DIF levels, ranked the same way they did in the baseline no DIF condition. Quartiles yielded an overall higher average classification rate of 80.09%, followed by deciles (79.06%), followed by four-adjacent (78.82%), followed by two-adjacent (78.45%). The differences among matching types were most evident in the smaller sample size conditions. Given that B-level items had the lowest classification rates, it is worthwhile to look further into the performance of matching types for B-level items. Quartiles still yielded the best classification rates for B-level items (29.29% average across sample sizes), closely followed by deciles (27.81% average across sample sizes), four-adjacent (27.23% average across sample sizes) and two-adjacent (27.17% average across sample sizes).

Discussion

The results of this study provide empirical demonstration of performance of both statistical and practical significance indicators employed in the MH technique for detecting DIF.

With regard to matching type, it appears that quartiles outperform others, but differences are not major. As for sample size, larger sample sizes yield better classification rates, with asymmetrical design giving an extra disadvantage, especially when total sample size is small. As a result of this study, a combination of quartile matching type and larger total sample size is optimal. In addition, it was found that items with B-level ETS DIF are the most problematic to detect using MH, across sample sizes and matching types.

Other conditions that have not been manipulated in the current study but might have an effect on the MH performance: test length, item difficulty, percentage of items with DIF, focal group proficiency level, use of purification techniques, direction of DIF, and the effect of outliers and non-normal distributions in small sample sizes. Future simulation studies need to manipulate these factors in order to fully explore the MH performance on detecting DIF, especially for "borderline" B-level items.

References

Camilli, G., & Smith, J. K. (1990). Comparison of the Mantel-Haenszel test with a randomized and a jackknife test for detecting biased items. *Journal of Educational Statistics*, *15*, 53-67.

Donoghue, J. R. & Allen, N. A. (1993). Thin versus Thick Matching in the Mantel-Haenszel Procedure for Detecting DIF. *Journal of Educational Statistics,18* (2), 131-154.

Dorans, N.J. & P.W. Holland. (1993). DIF Detection and Description: Mantel-Haenszel and Standardization in P.W. Holland & H. Wainer (Eds.). *Differential Item Functioning* (pp. 35-66) Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

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

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). New Jersey: Lawrence Erlbaum Associates, Inc.

No Child Left Behind Act of 2001, 20 U.S.C.S. Section 6301 et seq. (2002). Retrieved March 2nd, 2011, from http://www.ed.gov/nclb/overview/importance/difference/index.html

Paek, I. & Guo, H. (2011). Accuracy of DIF estimates and power in unbalanced designs using the Mantel-Haenszel DIF detection procedure. *Applied Psychological Measurement, 35*(7), 518-535.

SAS Institute Inc. 2008. SAS/STAT® 9.2 User’s Guide. Cary, NC: SAS Institute Inc.

Scheuneman, J. (1979). A method for assessing bias in test items. *Journal of Educational Measurement, 16*(3), 143-152.

Schmitt, A., Holland, P., & Dorans, N. (1993). Evaluating hypotheses about differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 281-316). Hillsdale, NJ: Erlbaum

Schultz, E. M., Perlman, C. P., Rice, W. K., & Wright, B. D. (1989). Empirical comparison of Rasch and Mantel-Haenszel procedures. Presented at the annual meeting of the American Educational Research Association, Boston, MA.

Schulz, E. M., Perlman, C., Rice, W. K., & Wright, B. D. (1996). An empirical comparison of Rasch and Mantel-Haenszel procedures for assessing differential item functioning. In G. Engelhard, Jr. & M. Wilson (Eds.), *Objective measurement: Theory into practice* (Vol. 3, pp. 65-82). Norwood, NJ: Ablex.

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 W.P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337-347). Hillsdale, NJ: Lawrence Erlbaum.

Zwick, R. (1991). Effects of item order and context on estimation of NAEP reading proficiency. *Educational Measurement: Issues and Practice, 10*(3), 10-16.

Zwick, R., & Ercikan, K. (1989). Analysis of differential item functioning in the NAEP history assessment. Journal of Educational Measurement, 26, 44-66.

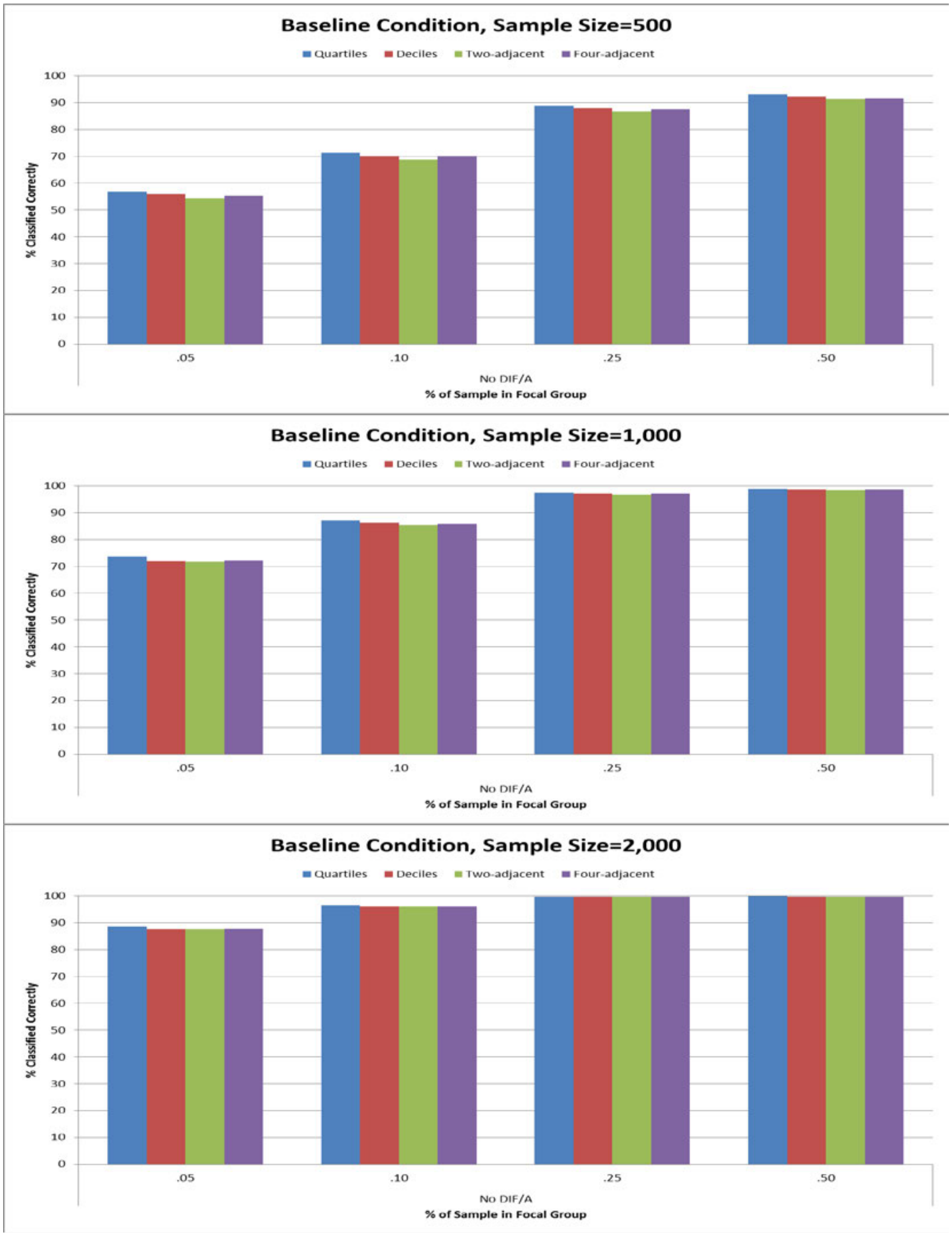


Figure 1. MH classification accuracy for baseline condition by sample size, percentage of sample in focal group, and matching type conditions (based on 100 replications).

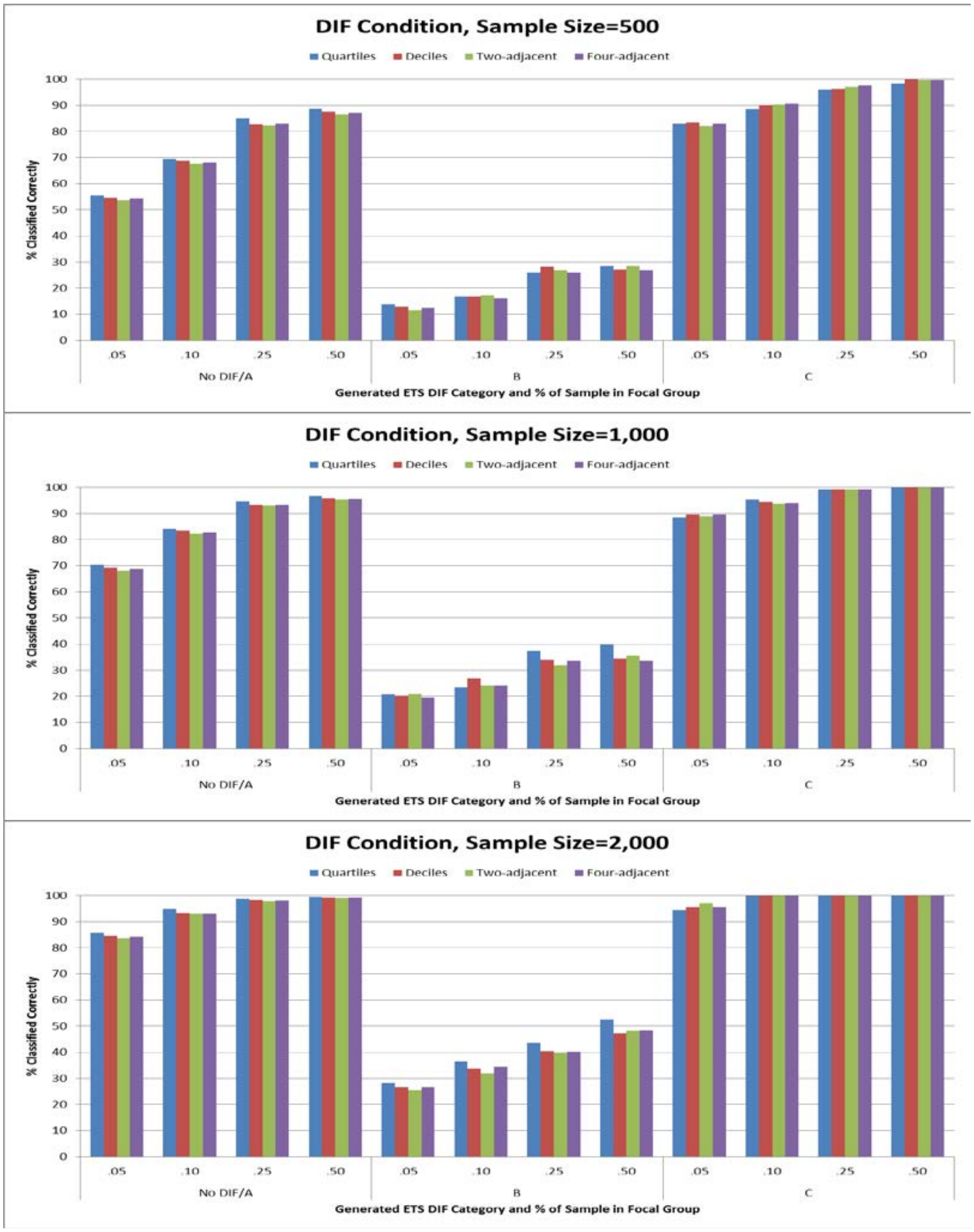


Figure 2. MH classification accuracy for the dataset with DIF by sample size, percentage of sample in focal group, and matching type conditions (based on 100 replications).

Appendix A

Descriptive Statistics of Estimated Δ_{MH} and True Positive Rate for MH Test of DIF: Baseline Condition (62 items)

Condition	Mean	Lower CI	Upper CI	True Positive Rate	A (%)	B (%)	C (%)
Sample Size = 500							
5% in Focal Group							
Quartiles	-0.040	-2.595	2.515	56.84	56.71	18.82	24.34
Deciles	-0.063	-2.703	2.577	55.95	55.82	18.00	26.05
Two-adjacent	-0.061	-2.787	2.665	54.29	54.15	18.52	27.19
Four-adjacent	-0.054	-2.721	2.613	55.27	55.15	18.34	26.39
10% in Focal Group							
Quartiles	-0.007	-1.836	1.822	71.22	71.20	17.13	11.66
Deciles	-0.027	-1.906	1.852	70.10	70.08	17.06	12.84
Two-adjacent	-0.033	-1.969	1.903	68.83	68.81	17.67	13.50
Four-adjacent	-0.020	-1.914	1.873	70.13	70.12	17.30	12.56
25% in Focal Group							
Quartiles	-0.013	-1.259	1.233	88.87	88.87	9.29	1.84
Deciles	-0.011	-1.290	1.269	88.06	88.06	9.76	2.18
Two-adjacent	-0.013	-1.327	1.302	86.71	86.71	10.81	2.48
Four-adjacent	-0.012	-1.301	1.277	87.48	87.48	10.53	1.98
50% in Focal Group							
Quartiles	-0.002	-1.078	1.074	93.05	93.05	6.24	0.71
Deciles	0.000	-1.104	1.104	92.18	92.18	7.03	0.79
Two-adjacent	0.004	-1.132	1.140	91.39	91.39	7.61	1.00
Four-adjacent	0.001	-1.111	1.114	91.74	91.74	7.40	0.85
Sample Size = 1,000							
5% in Focal Group							
Quartiles	-0.050	-1.823	1.724	73.53	73.52	15.77	10.69
Deciles	-0.036	-1.850	1.778	72.13	72.11	16.65	11.23
Two-adjacent	-0.030	-1.871	1.811	71.73	71.71	16.50	11.77
Four-adjacent	-0.031	-1.854	1.791	72.19	72.18	16.45	11.35
10% in Focal Group							
Quartiles	-0.017	-1.287	1.253	87.08	87.08	10.16	2.76
Deciles	-0.018	-1.318	1.281	86.21	86.21	10.82	2.97
Two-adjacent	-0.014	-1.333	1.304	85.40	85.40	11.48	3.11
Four-adjacent	-0.015	-1.322	1.291	85.87	85.87	11.13	3.00
25% in Focal Group							
Quartiles	0.006	-0.868	0.880	97.53	97.53	2.39	0.08
Deciles	-0.005	-0.899	0.889	97.11	97.11	2.74	0.15
Two-adjacent	-0.007	-0.914	0.901	96.69	96.69	3.06	0.24
Four-adjacent	-0.005	-0.904	0.894	97.02	97.02	2.82	0.16
50% in Focal Group							
Quartiles	0.000	-0.756	0.757	98.90	98.90	1.10	0.00
Deciles	0.002	-0.771	0.776	98.63	98.63	1.35	0.02

Appendix A

Descriptive Statistics of Estimated Δ_{MH} and True Positive Rate for MH Test of DIF: Baseline Condition (62 items)

Condition	Mean	Lower CI	Upper CI	True Positive Rate	A (%)	B (%)	C (%)
Two-adjacent	-0.004	-0.789	0.782	98.37	98.37	1.61	0.02
Four-adjacent	-0.003	-0.780	0.774	98.58	98.58	1.39	0.03
Sample Size = 2,000							
5% in Focal Group							
Quartiles	-0.018	-1.253	1.218	88.68	88.68	9.23	2.10
Deciles	-0.025	-1.287	1.236	87.58	87.58	10.00	2.42
Two-adjacent	-0.013	-1.286	1.260	87.68	87.68	9.73	2.60
Four-adjacent	-0.018	-1.283	1.247	87.84	87.84	9.73	2.44
10% in Focal Group							
Quartiles	0.012	-0.881	0.905	96.55	96.55	3.29	0.16
Deciles	0.002	-0.909	0.914	96.16	96.16	3.58	0.26
Two-adjacent	-0.004	-0.925	0.917	96.05	96.05	3.71	0.24
Four-adjacent	-0.005	-0.921	0.911	96.05	96.05	3.69	0.26
25% in Focal Group							
Quartiles	0.003	-0.612	0.619	99.82	99.82	0.18	0.00
Deciles	0.001	-0.627	0.630	99.82	99.82	0.18	0.00
Two-adjacent	-0.001	-0.636	0.633	99.71	99.71	0.29	0.00
Four-adjacent	-0.003	-0.633	0.628	99.73	99.73	0.27	0.00
50% in Focal Group							
Quartiles	-0.007	-0.540	0.527	99.90	99.90	0.10	0.00
Deciles	0.002	-0.543	0.546	99.87	99.87	0.13	0.00
Two-adjacent	0.001	-0.549	0.551	99.84	99.84	0.16	0.00
Four-adjacent	-0.001	-0.547	0.546	99.87	99.87	0.13	0.00

Note. True positive rate is the proportion of replications correctly classified as A (i.e., no DIF). CI – 95% Confidence Interval.

Appendix B

Descriptive Statistics of Estimated Δ_{MH} and True Positive Rate for MH Test of DIF: Items Generated without DIF (52 items)

Condition	Mean	Lower CI	Upper CI	True Positive Rate	A (%)	B (%)	C (%)
No DIF Items							
Sample Size = 500							
5% in Focal Group							
Quartiles	0.196	-2.360	2.751	55.63	55.50	19.46	24.90
Deciles	0.238	-2.386	2.863	54.56	54.42	18.40	27.04
Two-adjacent	0.256	-2.446	2.957	53.87	53.71	18.12	28.02
Four-adjacent	0.241	-2.402	2.885	54.44	54.31	18.42	27.13
10% in Focal Group							
Quartiles	0.210	-1.624	2.044	69.79	69.79	17.81	12.40
Deciles	0.240	-1.641	2.121	69.17	69.17	17.37	13.46
Two-adjacent	0.261	-1.675	2.197	67.92	67.92	17.73	14.35
Four-adjacent	0.248	-1.649	2.146	68.52	68.52	17.52	13.96
25% in Focal Group							
Quartiles	0.232	-1.019	1.483	85.60	85.60	11.71	2.69
Deciles	0.289	-0.995	1.572	82.98	82.98	12.88	4.13
Two-adjacent	0.284	-1.035	1.602	82.62	82.62	13.13	4.25
Four-adjacent	0.281	-1.012	1.574	83.23	83.23	12.81	3.96
50% in Focal Group							
Quartiles	0.250	-0.825	1.324	89.44	89.44	9.02	1.54
Deciles	0.288	-0.814	1.391	88.04	88.04	9.96	2.00
Two-adjacent	0.291	-0.840	1.422	87.06	87.06	10.75	2.19
Four-adjacent	0.288	-0.822	1.399	87.58	87.58	10.29	2.13
Sample Size = 1,000							
5% in Focal Group							
Quartiles	0.228	-1.541	1.997	71.02	71.02	17.40	11.58
Deciles	0.243	-1.564	2.051	69.71	69.71	18.04	12.25
Two-adjacent	0.263	-1.573	2.099	68.56	68.56	18.13	13.31
Four-adjacent	0.250	-1.566	2.066	69.13	69.13	18.33	12.54
10% in Focal Group							
Quartiles	0.232	-1.040	1.505	84.75	84.75	12.13	3.12
Deciles	0.271	-1.028	1.570	83.85	83.85	12.40	3.75
Two-adjacent	0.279	-1.040	1.599	82.77	82.77	13.10	4.13
Four-adjacent	0.275	-1.031	1.581	83.33	83.33	12.77	3.90
25% in Focal Group							
Quartiles	0.249	-0.626	1.124	95.08	95.08	4.58	0.35
Deciles	0.279	-0.616	1.173	93.83	93.83	5.75	0.42
Two-adjacent	0.286	-0.621	1.194	93.50	93.50	6.04	0.46
Four-adjacent	0.284	-0.615	1.183	93.83	93.83	5.69	0.48
50% in Focal Group							

Appendix B

Descriptive Statistics of Estimated Δ_{MH} and True Positive Rate for MH Test of DIF: Items Generated without DIF (52 items)

Condition	Mean	Lower CI	Upper CI	True Positive Rate	A (%)	B (%)	C (%)
Quartiles	0.244	-0.513	1.001	97.00	97.00	2.92	0.08
Deciles	0.284	-0.491	1.058	96.17	96.17	3.69	0.13
Two-adjacent	0.292	-0.495	1.078	95.79	95.79	3.98	0.23
Four-adjacent	0.287	-0.491	1.066	96.00	96.00	3.81	0.19
Sample Size = 2,000							
5% in Focal Group							
Quartiles	0.229	-1.009	1.467	86.21	86.21	11.08	2.71
Deciles	0.262	-1.000	1.524	84.81	84.81	12.08	3.12
Two-adjacent	0.268	-1.006	1.542	83.87	83.87	12.75	3.38
Four-adjacent	0.265	-1.002	1.531	84.48	84.48	12.19	3.33
10% in Focal Group							
Quartiles	0.232	-0.664	1.127	95.35	95.35	4.52	0.13
Deciles	0.277	-0.636	1.190	93.83	93.83	5.92	0.25
Two-adjacent	0.280	-0.641	1.201	93.83	93.83	5.87	0.31
Four-adjacent	0.279	-0.637	1.195	93.62	93.62	6.08	0.31
25% in Focal Group							
Quartiles	0.235	-0.383	0.852	99.13	99.13	0.85	0.02
Deciles	0.273	-0.357	0.902	98.71	98.71	1.29	0.00
Two-adjacent	0.283	-0.352	0.918	98.38	98.38	1.62	0.00
Four-adjacent	0.280	-0.352	0.912	98.56	98.56	1.44	0.00
50% in Focal Group							
Quartiles	0.245	-0.290	0.780	99.73	99.73	0.27	0.00
Deciles	0.274	-0.272	0.821	99.48	99.48	0.52	0.00
Two-adjacent	0.284	-0.268	0.835	99.37	99.37	0.63	0.00
Four-adjacent	0.282	-0.266	0.830	99.46	99.46	0.54	0.00

Note. True positive rate is the proportion of replications having the same ETS classification and direction as how the items were generated. CI – 95% Confidence Interval.

Appendix C

Descriptive Statistics of Estimated Δ_{MH} and True Positive Rate for MH Test of DIF: DIF Condition Items Generated as A Level (Items: 3, 28, 52)

Condition	Mean	Lower CI	Upper CI	True Positive Rate	A (%)	B (%)	C (%)
A Generated Items							
Sample Size = 500							
5% in Focal Group							
Quartiles	-0.679	-3.346	1.988	54.33	54.00	13.00	32.67
Deciles	-0.669	-3.426	2.088	53.33	53.00	14.00	32.67
Two-adjacent	-0.646	-3.485	2.193	50.00	49.67	16.33	33.67
Four-adjacent	-0.675	-3.451	2.101	51.67	51.33	15.00	33.33
10% in Focal Group							
Quartiles	-0.572	-2.471	1.328	63.33	63.33	19.33	17.33
Deciles	-0.579	-2.530	1.373	61.00	61.00	21.33	17.67
Two-adjacent	-0.561	-2.569	1.448	62.33	62.33	18.33	19.33
Four-adjacent	-0.577	-2.547	1.394	59.67	59.67	22.67	17.67
25% in Focal Group							
Quartiles	-0.536	-1.802	0.730	76.00	76.00	17.00	7.00
Deciles	-0.515	-1.815	0.786	78.00	78.00	15.33	6.67
Two-adjacent	-0.539	-1.875	0.796	78.00	78.00	12.33	9.67
Four-adjacent	-0.532	-1.842	0.778	78.33	78.33	12.33	9.33
50% in Focal Group							
Quartiles	-0.511	-1.599	0.577	78.67	78.67	18.33	3.00
Deciles	-0.505	-1.622	0.613	78.33	78.33	17.00	4.67
Two-adjacent	-0.496	-1.645	0.653	77.67	77.67	16.00	6.33
Four-adjacent	-0.504	-1.631	0.623	78.67	78.67	16.67	4.67
Sample Size = 1,000							
5% in Focal Group							
Quartiles	-0.602	-2.450	1.247	61.00	61.00	22.33	16.67
Deciles	-0.617	-2.517	1.283	60.67	60.67	21.33	18.00
Two-adjacent	-0.603	-2.532	1.325	61.33	61.33	20.67	18.00
Four-adjacent	-0.610	-2.519	1.298	61.33	61.33	20.67	18.00
10% in Focal Group							
Quartiles	-0.567	-1.882	0.748	73.67	73.67	17.00	9.33
Deciles	-0.554	-1.899	0.791	74.33	74.33	17.00	8.67
Two-adjacent	-0.549	-1.915	0.817	74.00	74.00	16.33	9.67
Four-adjacent	-0.551	-1.903	0.801	74.00	74.00	16.33	9.67
25% in Focal Group							
Quartiles	-0.510	-1.405	0.385	86.00	86.00	12.00	2.00
Deciles	-0.510	-1.426	0.406	85.33	85.33	12.67	2.00
Two-adjacent	-0.500	-1.429	0.429	87.00	87.00	10.33	2.67
Four-adjacent	-0.504	-1.425	0.417	86.33	86.33	12.00	1.67
50% in Focal Group							

Appendix C

Descriptive Statistics of Estimated Δ_{MH} and True Positive Rate for MH Test of DIF: DIF Condition Items Generated as A Level (Items: 3, 28, 52)

Quartiles	-0.520	-1.291	0.252	89.33	89.33	10.33	0.33
Deciles	-0.511	-1.301	0.279	88.67	88.67	11.00	0.33
Two-adjacent	-0.511	-1.313	0.292	89.67	89.67	9.67	0.67
Four-adjacent	-0.511	-1.306	0.283	89.33	89.33	10.33	0.33
Sample Size = 2,000							
5% in Focal Group							
Quartiles	-0.421	-1.681	0.839	78.33	78.33	15.00	6.67
Deciles	-0.415	-1.702	0.873	81.33	81.33	12.33	6.33
Two-adjacent	-0.407	-1.707	0.892	79.67	79.67	13.67	6.67
Four-adjacent	-0.419	-1.711	0.874	80.67	80.67	12.67	6.67
10% in Focal Group							
Quartiles	-0.551	-1.470	0.368	84.00	84.00	14.67	1.33
Deciles	-0.533	-1.471	0.405	85.00	85.00	13.33	1.67
Two-adjacent	-0.535	-1.482	0.412	82.67	82.67	14.67	2.67
Four-adjacent	-0.536	-1.478	0.406	83.33	83.33	15.00	1.67
25% in Focal Group							
Quartiles	-0.545	-1.177	0.087	91.33	91.33	8.67	0.00
Deciles	-0.534	-1.179	0.111	91.00	91.00	9.00	0.00
Two-adjacent	-0.527	-1.179	0.124	91.00	91.00	9.00	0.00
Four-adjacent	-0.530	-1.178	0.117	91.33	91.33	8.67	0.00
50% in Focal Group							
Quartiles	-0.514	-1.059	0.031	96.67	96.67	3.33	0.00
Deciles	-0.512	-1.069	0.044	96.33	96.33	3.67	0.00
Two-adjacent	-0.511	-1.073	0.051	96.33	96.33	3.67	0.00
Four-adjacent	-0.509	-1.067	0.050	96.00	96.00	4.00	0.00

Note. True positive rate is the proportion of replications having the same ETS classification and direction as how the items were generated. CI – 95% Confidence Interval.

Appendix D

Descriptive Statistics of Estimated Δ_{MH} and True Positive Rate for MH Test of DIF: DIF Condition Items Generated as B (Items: 21, 26, 43, 50)

Condition	Mean	Lower CI	Upper CI	True Positive Rate	A (%)	B (%)	C (%)
B Generated Items							
Sample Size = 500							
5% in Focal Group							
Quartiles	-1.694	-4.501	1.113	13.75	30.25	15.25	54.00
Deciles	-1.732	-4.652	1.189	13.00	28.75	15.00	55.75
Two-adjacent	-1.808	-4.863	1.247	11.50	29.50	12.75	57.25
Four-adjacent	-1.765	-4.734	1.205	12.50	28.25	14.75	56.50
10% in Focal Group							
Quartiles	-1.588	-3.615	0.439	16.75	30.50	17.25	52.00
Deciles	-1.638	-3.738	0.462	16.75	29.50	17.00	53.25
Two-adjacent	-1.642	-3.805	0.521	17.25	29.25	18.00	52.50
Four-adjacent	-1.646	-3.770	0.478	16.00	30.00	16.50	53.25
25% in Focal Group							
Quartiles	-1.580	-2.922	-0.238	26.00	18.75	26.00	55.25
Deciles	-1.588	-2.972	-0.205	28.25	18.00	28.25	53.75
Two-adjacent	-1.595	-3.019	-0.170	26.75	18.00	26.75	55.25
Four-adjacent	-1.601	-2.998	-0.204	26.00	17.50	26.00	56.50
50% in Focal Group							
Quartiles	-1.517	-2.642	-0.391	28.50	19.25	28.50	52.25
Deciles	-1.548	-2.705	-0.391	27.00	19.00	27.00	54.00
Two-adjacent	-1.568	-2.761	-0.375	28.50	17.50	28.50	54.00
Four-adjacent	-1.557	-2.724	-0.389	26.75	19.00	26.75	54.25
Sample Size = 1,000							
5% in Focal Group							
Quartiles	-1.592	-3.541	0.357	20.75	26.25	21.75	52.00
Deciles	-1.652	-3.664	0.360	20.25	24.75	21.00	54.25
Two-adjacent	-1.642	-3.684	0.400	21.00	25.00	21.75	53.25
Four-adjacent	-1.653	-3.674	0.369	19.50	24.75	20.50	54.75
10% in Focal Group							
Quartiles	-1.534	-2.911	-0.156	23.50	22.00	23.50	54.50
Deciles	-1.563	-2.977	-0.149	26.75	20.50	26.75	52.75
Two-adjacent	-1.574	-3.015	-0.132	24.00	21.50	24.00	54.50
Four-adjacent	-1.565	-2.990	-0.141	24.00	22.25	24.00	53.75
25% in Focal Group							
Quartiles	-1.487	-2.416	-0.557	37.25	15.25	37.25	47.50
Deciles	-1.527	-2.482	-0.572	34.00	13.50	34.00	52.50
Two-adjacent	-1.531	-2.501	-0.560	32.00	15.00	32.00	53.00
Four-adjacent	-1.529	-2.490	-0.569	33.50	14.00	33.50	52.50
50% in Focal Group							
Quartiles	-1.509	-2.298	-0.720	39.75	10.00	39.75	50.25
Deciles	-1.539	-2.347	-0.730	34.50	10.50	34.50	55.00

Descriptive Statistics of Estimated Δ_{MH} and True Positive Rate for MH Test of DIF: DIF Condition Items Generated as B (Items: 21, 26, 43, 50)

Two-adjacent	-1.544	-2.367	-0.720	35.50	11.00	35.50	53.50
Four-adjacent	-1.541	-2.355	-0.726	33.50	10.25	33.50	56.25
Sample Size = 2,000							
5% in Focal Group							
Quartiles	-1.591	-2.963	-0.219	28.25	18.50	28.50	53.00
Deciles	-1.625	-3.035	-0.214	26.50	19.00	26.50	54.50
Two-adjacent	-1.638	-3.066	-0.210	25.50	19.25	25.50	55.25
Four-adjacent	-1.639	-3.058	-0.220	26.50	18.00	26.50	55.50
10% in Focal Group							
Quartiles	-1.496	-2.464	-0.529	36.50	14.25	36.50	49.25
Deciles	-1.517	-2.509	-0.524	33.75	15.75	33.75	50.50
Two-adjacent	-1.527	-2.531	-0.523	32.00	17.00	32.00	51.00
Four-adjacent	-1.523	-2.521	-0.525	34.50	15.75	34.50	49.75
25% in Focal Group							
Quartiles	-1.506	-2.166	-0.846	43.50	7.25	43.50	49.25
Deciles	-1.536	-2.212	-0.860	40.50	6.00	40.50	53.50
Two-adjacent	-1.539	-2.223	-0.855	39.50	5.75	39.50	54.75
Four-adjacent	-1.536	-2.215	-0.857	40.25	6.00	40.25	53.75
50% in Focal Group							
Quartiles	-1.480	-2.041	-0.918	52.50	3.50	52.50	44.00
Deciles	-1.518	-2.093	-0.942	47.25	2.50	47.25	50.25
Two-adjacent	-1.526	-2.108	-0.944	48.25	2.25	48.25	49.50
Four-adjacent	-1.519	-2.097	-0.941	48.50	2.50	48.50	49.00

Note. True positive rate is the proportion of replications having the same ETS classification and direction as how the items were generated. CI – 95% Confidence Interval.

Appendix E

Descriptive Statistics of Estimated Δ_{MH} and True Positive Rate for MH Test of DIF: DIF Condition Items Generated as C (Items: 6, 23, 37)

Condition	Mean	Lower CI	Upper CI	True Positive Rate	A (%)	B (%)	C (%)
C Generated Items							
Sample Size = 500							
5% in Focal Group							
Quartiles	-3.043	-6.205	0.118	83.00	8.00	5.33	83.33
Deciles	-3.195	-6.548	0.157	83.33	7.67	5.33	83.67
Two-adjacent	-3.241	-6.740	0.259	82.00	9.00	5.33	82.33
Four-adjacent	-3.214	-6.610	0.182	83.00	7.67	5.67	83.33
10% in Focal Group							
Quartiles	-2.884	-5.126	-0.642	88.67	4.00	6.67	88.67
Deciles	-2.987	-5.326	-0.648	90.00	3.33	6.00	90.00
Two-adjacent	-2.995	-5.414	-0.577	90.33	3.67	5.33	90.33
Four-adjacent	-3.022	-5.407	-0.638	90.67	2.67	6.00	90.67
25% in Focal Group							
Quartiles	-2.863	-4.345	-1.381	96.00	0.00	4.00	96.00
Deciles	-2.936	-4.478	-1.394	96.33	0.00	3.67	96.33
Two-adjacent	-2.984	-4.587	-1.381	97.00	0.00	3.00	97.00
Four-adjacent	-2.981	-4.546	-1.416	97.67	0.00	2.33	97.67
50% in Focal Group							
Quartiles	-2.771	-3.970	-1.573	98.33	0.33	1.33	98.33
Deciles	-2.862	-4.103	-1.621	100.00	0.00	0.00	100.00
Two-adjacent	-2.877	-4.162	-1.592	99.67	0.00	0.33	99.67
Four-adjacent	-2.887	-4.144	-1.631	99.67	0.00	0.33	99.67
Sample Size = 1,000							
5% in Focal Group							
Quartiles	-2.892	-5.098	-0.687	88.33	5.33	6.33	88.33
Deciles	-3.005	-5.297	-0.713	89.67	3.33	7.00	89.67
Two-adjacent	-3.024	-5.367	-0.680	89.00	4.00	7.00	89.00
Four-adjacent	-3.033	-5.345	-0.720	89.67	3.67	6.67	89.67
10% in Focal Group							
Quartiles	-2.773	-4.316	-1.230	95.33	1.67	3.00	95.33
Deciles	-2.859	-4.458	-1.261	94.33	1.33	4.33	94.33
Two-adjacent	-2.875	-4.508	-1.241	93.67	2.00	4.33	93.67
Four-adjacent	-2.877	-4.490	-1.263	94.00	1.33	4.67	94.00
25% in Focal Group							
Quartiles	-2.802	-3.837	-1.766	99.33	0.00	0.67	99.33
Deciles	-2.910	-3.983	-1.836	99.33	0.00	0.67	99.33
Two-adjacent	-2.935	-4.032	-1.839	99.33	0.00	0.67	99.33
Four-adjacent	-2.928	-4.011	-1.845	99.33	0.00	0.67	99.33
50% in Focal Group							
Quartiles	-2.801	-3.651	-1.950	100.00	0.00	0.00	100.00
Deciles	-2.897	-3.776	-2.018	100.00	0.00	0.00	100.00
Two-adjacent	-2.925	-3.824	-2.027	100.00	0.00	0.00	100.00
Four-adjacent	-2.912	-3.798	-2.026	100.00	0.00	0.00	100.00
Sample Size = 2,000							
5% in Focal Group							
Quartiles	-2.822	-4.375	-1.268	94.33	0.67	5.00	94.33

Appendix E

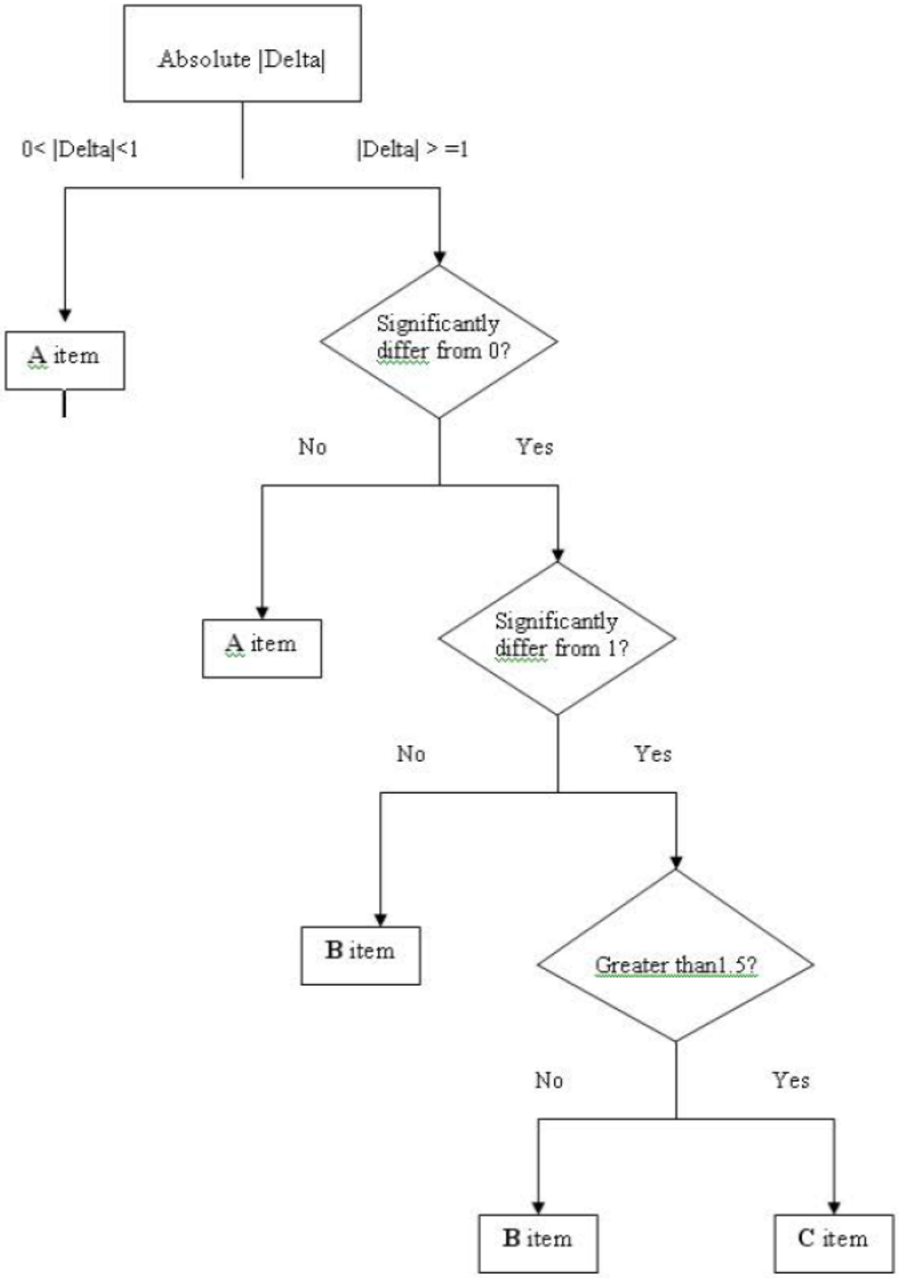
Descriptive Statistics of Estimated Δ_{MH} and True Positive Rate for MH Test of DIF: DIF Condition Items Generated as C (Items: 6, 23, 37)

Deciles	-2.925	-4.539	-1.311	95.67	0.67	3.67	95.67
Two-adjacent	-2.949	-4.587	-1.312	97.00	0.67	2.33	97.00
Four-adjacent	-2.942	-4.567	-1.317	95.67	0.33	4.00	95.67
10% in Focal Group							
Quartiles	-2.793	-3.859	-1.728	100.00	0.00	0.00	100.00
Deciles	-2.876	-3.978	-1.774	100.00	0.00	0.00	100.00
Two-adjacent	-2.909	-4.029	-1.789	100.00	0.00	0.00	100.00
Four-adjacent	-2.887	-3.995	-1.778	100.00	0.00	0.00	100.00
25% in Focal Group							
Quartiles	-2.772	-3.494	-2.050	100.00	0.00	0.00	100.00
Deciles	-2.858	-3.603	-2.113	100.00	0.00	0.00	100.00
Two-adjacent	-2.879	-3.636	-2.122	100.00	0.00	0.00	100.00
Four-adjacent	-2.868	-3.619	-2.118	100.00	0.00	0.00	100.00
50% in Focal Group							
Quartiles	-2.743	-3.346	-2.140	100.00	0.00	0.00	100.00
Deciles	-2.840	-3.462	-2.219	100.00	0.00	0.00	100.00
Two-adjacent	-2.860	-3.491	-2.230	100.00	0.00	0.00	100.00
Four-adjacent	-2.850	-3.476	-2.225	100.00	0.00	0.00	100.00

Note. True positive rate is the proportion of replications having the same ETS classification as how the items were generated. CI – 95% Confidence Interval.

Appendix F

ETS DIF Classification Rules


Absolute |Delta|
0< |Delta|<1
|Delta| > =1
A item
Significantly differ from 0?
No
Yes
A item
Significantly differ from 1?
No
Yes
B item
Greater than1.5?
No
Yes
B item
C item