Research Report

Robustness of Weighted Differential Item Functioning (DIF) Analysis: The Case of Mantel–Haenszel DIF Statistics

ETS RR–21-12

Ru Lu
Hongwen Guo
Neil J. Dorans

December 2021

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Robustness of Weighted Differential Item Functioning (DIF) Analysis: The Case of Mantel–Haenszel DIF Statistics

Ru Lu, Hongwen Guo, & Neil J. Dorans

ETS, Princeton, NJ

Two families of analysis methods can be used for differential item functioning (DIF) analysis. One family is DIF analysis based on observed scores, such as the Mantel – Haenszel (MH) and the standardized proportion-correct metric for DIF procedures; the other is analysis based on latent ability, in which the statistic is a measure of departure from measurement invariance (DMI) for two studied groups. Previous research has shown, that DIF and DMI do not necessarily agree with each other. In practice, many operational testing programs use the MH DIF procedure to flag potential DIF items. Recently, weighted DIF statistics has been proposed, where weighted sum scores are used as the matching variable and the weights are the item discrimination parameters. It has been shown theoretically and analytically that, given the item parameters, weighted DIF statistics can close the gap between DIF and DMI. The current study investigates the robustness of using weighted DIF statistics empirically through simulations when item parameters have to be estimated from data.

Keywords DIF; measurement invariance; matching variable; IRT

doi:10.1002/ets2.12325

Many testing programs routinely conduct differential item functioning (DIF) analysis to check whether there are items on the tests that favor one group of test takers over another for group members of comparable proficiency. There are two major approaches for DIF analysis: One approach relies on observed test scores, and the other uses item response theory (IRT) models. Among DIF methods based on observed scores, the Mantel – Haenszel DIF (MH DIF) was introduced to the educational assessment field by Holland and Thayer (1988) to study how dichotomously scored items function between different groups and has been used in many testing programs. Another observed-score method is the standardized proportion-correct metric for DIF (P-DIF) by Dorans and Kulick (1986). For the IRT-based DIF approach (e.g., Thissen et al., 1988, 1993), psychometricians use either the differences in item parameters or areas between item response functions for the two studied groups of test takers as a measure of departure from measurement invariance (DMI). The IRT-based method requires that the selected IRT model fits the data and a larger sample for both groups (Clauser & Mazor, 1998; Holland & Thayer, 1988). However, the observed-score DIF methods can work with both small and large samples. The observed-score method contains implicit assumptions, too. For example, Camilli and Shepard (1994) pointed out MH DIF statistics are unbiased for multiple-choice items only when several assumptions are met, one of which is that the data fit the Rasch model. More research (e.g., Donoghue et al., 1993; Roussos et al., 1999) showed that when two- or three-parameter logistic (2PL or 3PL) models are fitted to the data, the MH DIF statistics differed from the latent-ability based DMI.

Recently, Guo and Dorans (2019, 2020) made a distinction between DIF and DMI, in which DIF is defined as the statistics that are based on observed scores to assess items' differential performance across different groups of test takers, and DMI is defined as the departure from measurement invariance based on the measured latent ability. Guo and Dorans (2019, 2020) proposed the use of weighted sum scores (where weights are equal to the estimated discrimination parameters) in the DIF context. Through analytic methods, they showed that the DIF indices of effect sizes are different from the DMI measure under the 2PL model in many situations when a simple sum score is used as the matching or conditioning variable in DIF analysis. However, they demonstrated that matching on a weighted sum score can significantly reduce the difference between DIF and DMI statistics if the 2PL model fits the data. To obtain weighted sum scores, item discrimination parameters are used as the item score weights, and thus the number of possible weighted sum scores can be very large, which in practice makes it infeasible to implement the weighted DIF method. Guo and Dorans (2019, 2020) proposed binning methods (e.g., naive, linear, and equipercentile) to facilitate the potential use of weighted DIF in practice.

Corresponding author: R. Lu, E-mail: rlu@ets.org

Assuming that the 2PL model is the appropriate model for item responses, Guo and Dorans (2019) analytically showed that DIF statistics based on the simple sum score (i.e., MH DIF and P-DIF) differed from the IRT-based DMI measure when the compared groups have different abilities; that is, DIF statistics based on the simple sum score are nonzero under the null hypothesis that DMI is zero. For uniform DIF scenarios, Guo and Dorans found that the most important influence on the differences between DIF and DMI is the ability difference between the focal and reference groups. A second important factor was the variation of the item discrimination parameters. When the studied item has a discrimination parameter close to the average discrimination of the test, DIF and DMI exhibit small differences. The item difficulty parameter also impacts the differences between DIF and DMI: When the item is relatively harder, the DIF size is slightly smaller than the corresponding DMI measure for a weaker focal group. In contrast, the weighted DIF size is zero under the null hypothesis that the DMI is zero. In the case of the DMI not being equal to zero, DIF statistics based on the weighted sum scores still produced a much smaller difference from DMI than DIF statistics based on the simple sum scores.

Findings in Guo and Dorans (2019, 2020) were population results, which used model assumptions and numerical calculations without real data. A number of previous simulation studies have shown that both MH DIF and P-DIF statistics are affected by additional factors in practice, such as test length and sample sizes (e.g., Zwick et al., 1997). It is reasonable to suspect that test length and sample sizes also impact weighted DIF statistics. Moreover, in practice, item parameters are not known. They need to be estimated with real data. Thus, the purpose of the current study is to investigate the robustness of weighted DIF statistics for several real-situation factors: sample size, test length, ability difference, and, most importantly, item parameter estimation error. In this study, we focused on the MH DIF statistics. In the next section, we briefly review the definitions of DIF and DMI used in Guo and Dorans (2019, 2020) and this study: the calculation of MH DIF statistics matched with the simple sum scores and weighted sum scores. Then we present the simulation design, the results, and a discussion of the application of weighted DIF statistics for operational use.

The MH DIF Statistics

This study focuses on the performance of MH DIF statistics. First, we define the terms DIF and DMI used in this study, and then we illustrate MH DIF statistics with simple sum scores as the matching variable and MH DIF statistics matched with the weighted sum scores, using three different binning methods. Note that the definitions or calculations are consistent with Guo and Dorans (2019, 2020). For more discussion of these definitions, please refer to Guo and Dorans (2019, 2020).

Definitions of DMI and DIF

For a test that has J dichotomously scored items, let Y stand for the score of item j ($y_j = 1$) when the response is correct and $y_j = 0$ when the response is wrong), let X represent the observed sum score, and let θ be the latent ability of a test taker measured on the test. Further assuming that the item responses follow the 2PL model, the probability of test takers obtaining a correct response to item j would be:

$$P\left(Y_j = 1|\theta\right) = \frac{\exp\left[D_0 a_j\left(\theta - b_j\right)\right]}{1 + \exp\left[D_0 a_j\left(\theta - b_j\right)\right]}, \tag{1}$$

where $D_0 = 1.7$, and a_j and b_j are the item discrimination and difficulty parameters (Lord, 1980). The DMI measure is defined as the log odds ratio matched on the latent ability, thus,

$$\text{DMI} = -2.35 \int_\theta \left\{ \frac{P_r(\theta)/Q_r(\theta)}{P_f(\theta)/Q_f(\theta)} \right\} \Psi_r(\theta)\, d\theta, \tag{2}$$

where $P_r(\theta) = P_r\left(Y_j = 1|\theta\right)$; $P_f(\theta) = P_f\left(Y_j = 1|\theta\right)$; they are the item response functions for a correct response in the reference and focal groups, respectively. Correspondingly, $Q_r(\theta) = 1 - P_r(\theta)$; $Q_f(\theta) = 1 - P_f(\theta)$ are the item response functions for an incorrect response in the reference and focal groups, respectively. $\Psi_r(\theta)$ is the density function of the latent ability for the reference group (Guo & Dorans, 2020; Zwick et al., 2000).

Similarly, DIF is defined as the expected value of the log odds ratio with respect to the observed-score distribution of the reference group, that is,

$$\mathrm{DIF} = -2.35 \sum_{x} \ln \left\{ \frac{P_r(x)/Q_r(x)}{P_f(x)/Q_f(x)} \right\} g_r(x), \tag{3}$$

where $g_r(x)$ is the probability distribution of X in the reference form. Note that the measure in Equation 3 is different than operational MH DIF statistics (Holland & Thayer, 1988), but both measures were designed to estimate the same DMI (refer to Guo & Dorans, 2020, for discussion).

Operational MH DIF Statistics With Simple Sum Scores

The simple sum score is the number correct score. It is obtained as

$$X = \sum_{j=1}^{J} Y_j, \tag{4}$$

where X is an integer value in the range of 0 to J for a test of J items. Test takers from reference and focal groups are assumed to be matched in their ability if they have the same X. In practice, to calculate the MH DIF statistics for item j, each level of X is treated as a stratum ($k = 0, 1, \ldots, J$). In stratum k, let R_{kr} and R_{kf} represent the number of test takers that have a correct response to item j in the reference and focal groups, respectively; W_{kr} and W_{kf} represent the number of test takers that have an incorrect response to item j in the reference and focal groups. Then, following Dorans and Kulick (1986) and Holland and Thayer (1988), MH DIF statistics are obtained as

$$\mathrm{MH} = -2.35 \ln \left(\frac{\sum_k R_{kr} W_{kf}/T_k}{\sum_k R_{kf} W_{kr}/T_k} \right), \tag{5}$$

where T_k is the total number of test takers in stratum k. The values of MH could be negative or positive. A negative value indicates that the item favors the reference group while a positive value indicates that the item favors the focal group. When there is no DIF, the MH DIF statistics are expected to be close to zero.

MH DIF Statistics with Weighted Sum Scores

The MH DIF statistics with weighted sum scores use the item discrimination parameters as weights, which, in practice, are never known. Thus, the item discrimination parameter estimates are used to calculate the weighted sum scores in this study. The weighted sum score, Z, is defined as:

$$Z = \sum_{j=1}^{J} \hat{a}_j Y_j, \tag{6}$$

where $\hat{a}_j$ is the item discrimination parameter estimate. In practice, the item discrimination parameter or its estimate could be any positive number, typically smaller than or equal to 2 (Harris, 1989). The number of possible values of weighted sum scores in Equation 6 is much larger than that of simple sum scores (i.e., $J + 1$ for the latter). Thus, it is not feasible to use exact matching as in the case of simple sum scores to calculate the MH DIF statistics. Instead, we apply the three binning methods (i.e., naive, linear, and equipercentile) proposed in Guo and Dorans (2019, 2020) to put weighted sum scores into a manageable number of bins. Test takers with the weighted sum scores that are within the same bin are considered matched in ability.

Binning Methods

In the naive binning method, the range of weighted sum scores is partitioned into $(J + 1)$ equal intervals (bins) for a test of length J. Test takers in the focal group whose weighted sum scores are in an interval (bin) are matched to those in the same interval in the reference group.

For the other two binning methods, linear and equipercentile, we applied techniques to find the boundaries between bins (Guo & Dorans, 2020; Kolen & Brennan, 2004). As in the naive binning method, test takers whose weighted sum

scores are within two adjacent boundaries were put into the same bin and considered matched in ability. More specifically, let $F_X(x)$ and $F_Z(z)$ stand for the cumulative score distributions of the simple sum score (X) and the weighted sum score (Z), respectively; μ_X and σ_X are the mean and standard deviation of the simple sum scores; and μ_Z and σ_Z are the mean and standard deviation of the weighted sum scores.

Under the linear binning method, for each possible score of X ($x = 0, 1, \ldots, J$), the corresponding Z score would be:

$$z_X = \mu_Z + \frac{\sigma_Z}{\sigma_X}\left(x - \mu_X\right), \tag{7}$$

The bin boundaries are defined as $\left(z_X - H\right)$ and $\left(z_X + H\right)$, where $H = \sigma_z / \left(2\sigma_X\right)$. Test takers whose weighted sum scores are larger than $\left(z_X - H\right)$ but smaller than or equal to $\left(z_X + H\right)$ are considered matched.

Under the equipercentile binning method, two test takers with the same percentile rank in their respective score distributions are considered to be equal. Thus, for each level of X, the corresponding weighted sum score would be

$$z_X = F_z^{-1}\left[F_X(x)\right], \tag{8}$$

Similarly, test takers whose weighted sum scores are in the same interval of $(z_X - H_X, z_X + H_X)$ are considered to be matched with $H_X = \left(z_X - z_{X-1}\right)/2$.

After matching test takers based upon their weighted sum scores, the calculation of the weighted MH DIF statistics are the same as in Equation 5.

The Design of the Simulation Study

It is necessary to conduct a simulation study to investigate the practical impact of MH DIF statistics using the weighted sum scores. The same test design in Guo and Dorans (2019, 2020) was applied in this study to bridge the pure analytic results and those obtained from real testing data. As stated in Guo and Dorans (2019), the design of the study was based upon previous studies (e.g., Camilli & Shepard, 1994; Zwick et al., 1997) and test information obtained from a large-scale testing program. Instead of using the observed item parameter estimates in operational settings, both the item discrimination and difficulty parameters are set to have three levels each in our simulations. Together, they give us nine items that span a realistic range of item statistics. The item parameters for the nine items are as follows:

- Item difficulty b_r for the reference group: b_r is set at -1, 0, or 1 to represent an easy, moderate, or hard item.
- Item discrimination a: a is set at .48, .60, or .75 in the 2PL model, which approximate the lower quartile, median, and the upper quartile of a log-normal distribution derived from a large-scale standardized test.

For the true DMI magnitude, we set three levels of item difficulty difference $d = b_f - b_r$ between the focal and reference groups; that is, d is set to $-.25$, 0, and .25, respectively. Crossing these three values with the above nine items, we have a 27-item hypothetical test form as shown in the second (item discrimination parameter) and third (item difficulty parameter) columns in Table 1.

Although we have a test of 27 items, we study only one item each time with the d specified in Table 1. For example, when we study the MH DIF statistics for the first item in Table 1, we suspect that the first item may function differently between the reference and focal groups; test takers' responses to this item are generated using the parameters of ($a = 0.48$, $b_r = -1$) for the reference group and ($a = 0.48$, $b_r = -0.75$) for the focal group. Each of the remaining items in the test is DMI-free; that is, the same item discrimination and difficulty parameters are used for Item 2 to Item 27 to generate responses for test takers in both the reference and focal groups. One can think of this as 27 versions of the same test, and each version focuses on a different studied item. Among them, nine versions of the test have no DMI; nine versions have positive DMI, which favors the focal group; and the remaining nine versions have negative DMI, which favors the reference group.

In addition to the three item-related factors (i.e., item discrimination, item difficulty, and item difficulty difference), we manipulated three other factors that matter in operational settings:

1. Test length: short and long. The number of items on the test was either $J = 27$ or $J = 54$. For the short test, the item discrimination and difficulty parameters are listed in Table 1. The long test is assumed to have two parallel short tests. Again, in each test version, we study only one potential DIF item, keeping the remaining items free of DMI.

Table 1 Item Parameters Used in the Simulation

Item ID	Item discrimination (a)	Item difficulty (b_r)	Item difficulty difference (d) ($d = b_r - b_f$)	DMI (θ)
1	0.48	−1	0.25	−0.4794
2	0.48	−1	0	0.0000
3	0.48	−1	−0.25	0.4794
4	0.48	0	0.25	−0.4794
5	0.48	0	0	0.0000
6	0.48	0	−0.25	0.4794
7	0.48	1	0.25	−0.4794
8	0.48	1	0	0.0000
9	0.48	1	−0.25	0.4794
10	0.6	−1	0.25	−0.5992
11	0.6	−1	0	0.0000
12	0.6	−1	−0.25	0.5993
13	0.6	0	0.25	−0.5992
14	0.6	0	0	0.0000
15	0.6	0	−0.25	0.5993
16	0.6	1	0.25	−0.5992
17	0.6	1	0	0.0000
18	0.6	1	−0.25	0.5993
19	0.75	−1	0.25	−0.7491
20	0.75	−1	0	0.0000
21	0.75	−1	−0.25	0.7491
22	0.75	0	0.25	−0.7491
23	0.75	0	0	0.0000
24	0.75	0	−0.25	0.7491
25	0.75	1	0.25	−0.7491
26	0.75	1	0	0.0000
27	0.75	1	−0.25	0.7491

Note. DMI = departure from measurement invariance.

Table 2 Sample Sizes Conditions

Group	Very small	Small	Moderate	Large	Very large
Focal	300	600	1,200	2,400	4,800
Reference	400	800	1,600	3,200	9,600

2. Ability difference: no or large difference. For the condition of no ability difference, the latent abilities of both reference and focal groups follow a standard normal distribution. For the condition of large ability difference, we set the ability difference between the focal and reference groups to be one standard deviation unit: The ability of the reference group follows a normal distribution with mean of 0.5 and standard deviation of 1; the ability of the focal group ability follows a normal distribution with a mean of −0.5 and standard deviation of 1.
3. Sample sizes. In practice, the sample sizes vary greatly for different testing programs and different groups of test takers. To make the simulation manageable, we studied five different pairs of sample sizes to mimic the reality (Zwick, 2012) and kept the ratio of sample sizes of the reference and focal groups constant (that is, the ratio was 4:3). The sample sizes conditions are summarized in Table 2.

The Analysis Procedure

The simulations were conducted in following steps:

1. Simulate test takers' responses. Using test length, ability difference, and sample sizes for the focal and reference groups, a DMI version of the test and item responses for the focal and reference groups was generated using the 2PL model.

2. Estimate item parameters. Based on the simulated response data, PARSCALE was used with the 2PL model to obtain the item parameter estimates. The combined group of test takers in both the reference and focal groups in particular were used in the calibration. Thus, for each set of responses generated in Step 1, there was only one set of item discrimination parameter estimates, as in an operational practice.
3. Calculate the MH DIF statistics. For each studied item in each simulation condition, we calculated the following four sets of MH DIF statistics: using the weighted sum scores as the matching variable (WSS MH DIF) with (a) naive binning, (b) linear binning, and (c) equipercentile binning, and (d) using the simple sum scores as the matching variable (SSS MH DIF).
4. Replicate steps. To account for sampling errors, we replicated 100 times Steps 1 – 3.

In total, 54,000 simulations (i.e., $27 \times 2 \times 2 \times 5 \times 100$) were conducted, and 216,000 MH DIF statistics (i.e., $54{,}000 \times 4$) were calculated.

Evaluation Criteria

To evaluate the accuracy and precision of the item discrimination parameter estimate under different simulation conditions, we calculated the departure of the estimated values from their true values, using bias and root mean square difference (RMSD). They are calculated as follows:

$$\text{Bias}_a = \frac{\sum_{i=1}^{100} \left(\hat{a} - a\right)}{100}, \tag{9}$$

$$\text{RMSD}_a = \sqrt{\frac{\sum_{i=1}^{100} \left(\hat{a} - a\right)^2}{100}}, \tag{10}$$

For each studied DMI item, we treated the population results of MH DIF in Guo and Dorans (2019, Table 2) as our DMI values and evaluated the departure of different MH DIF statistics (either matched with simple sum scores or weighted sum scores with a binning method) from the DMI, in terms of bias and RMSD. They are calculated as follows:

$$\text{Bias}_{\text{dif}} = \frac{\sum_{i=1}^{100} (DIF - DMI)}{100}, \tag{11}$$

$$\text{RMSD}_{\text{dif}} = \sqrt{\frac{\sum_{i=1}^{100} (DIF - DMI)^2}{100}}, \tag{12}$$

where DIF represents one of the four MH DIF statistics based on the simulated data, and DMI is the latent ability-based DMI population value obtained from Guo and Dorans (2019). The above analyses were conducted using SAS.

Simulation Results

We conducted simulations for all sample size conditions listed in Table 2. Because of space limitations, we present only the results from three sample sizes of the focal group: 300, 1,200, and 4,800. The results from the other sample sizes are available upon request. In addition, the intermediate results of the item discrimination parameter estimates are summarized in Appendix A; these estimates are consistent with previous IRT studies on item parameter estimation.

For the MH results, the order of presentation is based on test length and ability differences: The long test is followed by the short test, and no ability difference is followed by large ability difference, to emphasize how MH DIF and its weighted version become noticeably different. The values of bias and RMSD of all four MH DIF statistics are presented in Appendices B and C for the long and short test, respectively. The values of bias and RMSD of MH DIF statistics on the long test are presented in Tables B1 – B3 for the case of no ability difference and in Tables B4 – B6 for the case of large ability difference. The same statistics for the short test are presented in Appendix C in Tables C1 – C3 for the case of no ability difference and Tables C4 – C6 for the case of large ability difference.

Generally speaking, the values of the bias and RMSD of the three WSS MH DIF statistics are similar to each other in Tables B1 – C6. ANOVA results indicate no significant effect of different binning methods on WSS MH DIF statistics.

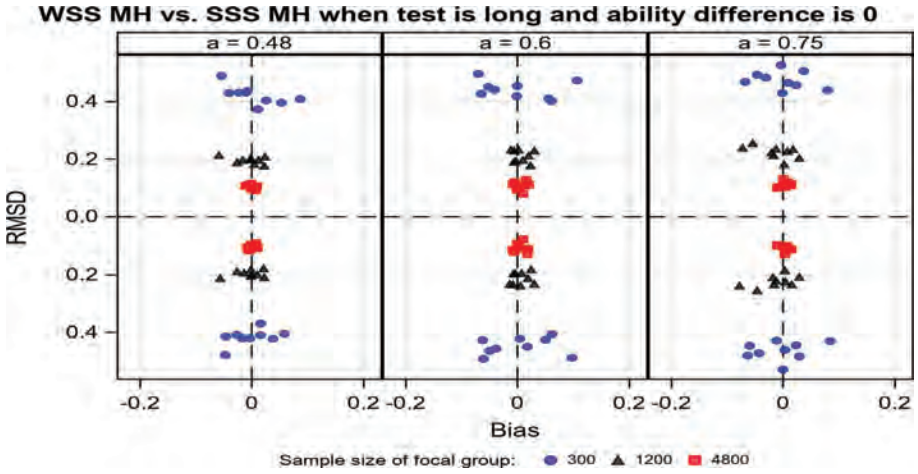

Figure 1 Bias and RMSD of MH DIF statistics for the long test with no ability difference. The figure contains three horizontal panels, one for each item discrimination parameter (a = .48, .60, or.75). Each panel has two vertical scatterplots: The upper scatterplot presents the bias and RMSD from the WSS MH-DIF analysis with naive binning; the lower scatterplot represents the bias and RMSD from the SSS MH-DIF analysis. *Note*. DIF = differential item functioning, MH = Mantel–Haenszel, RMSD = root mean square difference, SSS = simple sum scores, WSS = weighted sum scores.

Thus, in plotting the differences between WSS and SSS MH DIF statistics, we chose to use WSS MH DIF statistics with the naive binning method for its simplicity (and we recommend it be used in operational applications).

The Long Test

The Case of No Ability Difference

Figure 1 presents the bias and RMSD of MH DIF statistics for items in the case of the long test with no ability difference. In the plot, the vertical axis represents RMSD, and the horizontal axis represents bias. Horizontally, it has three panels, with each panel presenting items that have the same discrimination parameter: a = .48, .60, or .75. Vertically, each panel is divided into two scatterplots: The upper scatterplot contains only statistics for WSS MH DIF, and the lower scatterplot contains only statistics for SSS MH DIF. Note that the vertical axis scale is symmetric along the horizontal zero line for easy visual comparison. With such arrangements of the horizontal and vertical axis, we could present both the bias and RMSD of MH DIF statistics simultaneously. This demonstration was adapted from the item discrepancy indices plots in Dorans and Kulick (1986). In Figure 1, the blue circles represent results from a sample size of 300 in the focal group; the black triangles, from a sample size of 1,200 in the focal group; and the red squares, from a sample size of 4,800 in the focal group. Thus, in each scatterplot, we have either nine items in the column of a = .48 (i.e., Items 1–9 in Table 1), nine items in the column of a = .60 (i.e., Items 10–18 in Table 1), or nine items in the column of a = .75 (i.e., Items 19–27 in Table 1). Furthermore, each item appears three times with a different color and point symbol representing a different sample size condition.

In each scatterplot in Figure 1 (i.e., WSS MH DIF or SSS MH DIF), the conditions with larger sample sizes have a smaller magnitude of bias and RMSD; whereas the conditions with the smallest sample sizes have the largest magnitude of bias and RMSD and more variations than other sample size conditions. These findings indicate that, for the long test, the impact of sample sizes on the MH DIF statistics, either WSS or SSS, is the same: The estimates for WSS and SSS MH DIF statistics are more accurate with large sample sizes. If the sample sizes are small, the estimates for both WSS and SSS MH DIF statistics are less accurate and stable.

In Figure 1, the black triangles or the red squares in each panel almost mirror each other along the horizontal zero line, indicating that the bias and RMSD are generally similar for WSS and SSS MH DIF statistics for relatively large sample sizes (1,200 and above). Furthermore, across the three panels of different item discrimination parameters, the red squares are generally around the zero bias line and the zero RMSD line, indicating that with large sample sizes and no ability differences between the focal and reference groups, the discrepancy indices of WSS and SSS MH DIF statistics are similar.

The same pattern is observed when the three panels are grouped by the item difficulty parameters in the reference group (b_r), or the difference parameters (d) between the focal and reference groups (plots are available upon request).

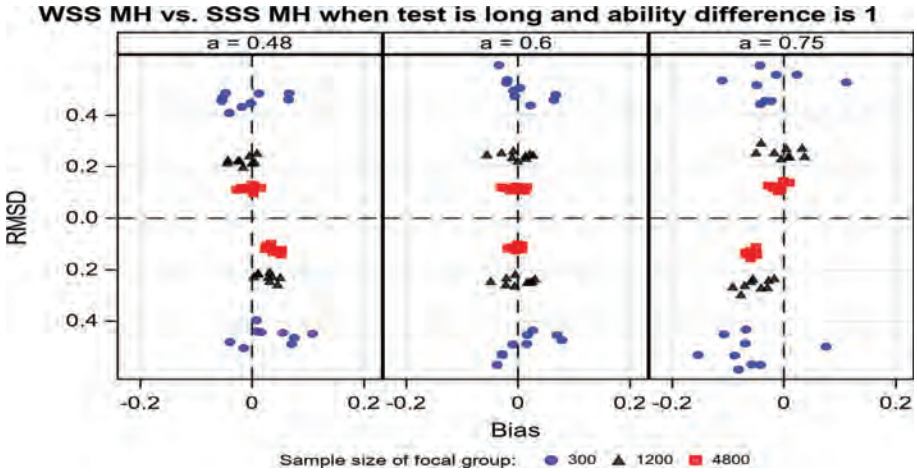

Figure 2 Bias and RMSD of MH DIF statistics for the long test with large ability difference. The figure contains three horizontal panels, one for each item discrimination parameter (a = .48, .60, or .75). Each panel has two vertical scatterplots: The upper scatterplot presents the bias and RMSD from the WSS MH DIF analysis with naive binning; the lower scatterplot represents the bias and RMSD from the SSS MH DIF analysis. *Note*. DIF = differential item functioning, MH = Mantel–Haenszel, RMSD = root mean square difference, SSS = simple sum scores, WSS = weighted sum scores.

That is, when the test is long and there are no group ability differences, there are hardly noticeable differences between WSS and SSS MH DIF statistics. Because of the extra steps of obtaining the item discrimination parameter estimates and the binning procedure associated with WSS MH DIF statistics, we see no advantage in choosing WSS over SSS MH DIF statistics in operational settings.

The Case of Large Ability Difference

Figure 2 presents the bias and RMSD in the case of the long test with large ability difference. Figure 2 follows the same format as Figure 1. Horizontally, there are three panels based on the item discrimination parameters. Vertically, there are two scatterplots in each panel in which the vertical scale is symmetric along the horizontal zero line: The upper panel presents the bias and RMSD of WSS MH DIF statistics, and the lower one presents the bias and RMSD of SSS MH DIF statistics.

In each scatterplot in Figure 2, the effect of sample sizes on either WSS or SSS MH DIF statistics is the same as the case of no ability difference: The larger the sample size, the smaller the magnitude of bias or RMSD for either WSS or SSS MH DIF statistics. However, in the panels for a = .48 or a = .75, we do not observe symmetric reflection along the vertical bias zero line, as in the case of no ability difference. Instead, even for the largest sample sizes in this study, the bias for SSS MH is positive for items with a = .48 (shown in the lower scatterplot of the left panel) and negative for items with a = .75 (shown in the lower scatterplot of the right panel). Only items with a = .60 (shown in the lower scatterplot in the middle panel) have close to zero bias. In contrast, WSS MH DIF statistics did not have positive or negative bias for items either with a = .48 or a = .75. All WSS MH DIF statistics are around the zero line. These results are consistent with Guo and Dorans (2019, 2020): If the item responses follow the 2PL model and there is large ability difference, SSS MH DIF statistics underestimate or overestimate DMI even with a long test, but WSS MH DIF statistics can approximate DMI closely. Furthermore, the positive or negative bias of SSS MH DIF statistics is not evident with different values of the item difficulty parameters or item difficulty difference parameters for items with a = .60. In all cases, WSS MH DIF statistics had close to zero biases.

The Short Test

The Case of No Ability Difference

Figure 3 presents the bias and RMSD for the MH DIF statistics in the case of the short test with no ability difference. The overall observations are similar to those in Figure 1. That is, when there is no ability difference between the focal and reference groups, there is no apparent difference between WSS and SSS MH DIF statistics for each set of item parameters:

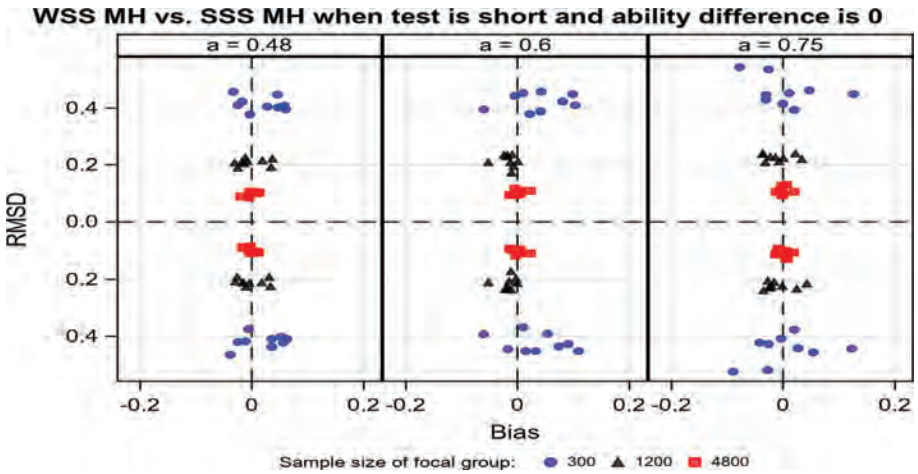

Figure 3 Bias and RMSD of MH DIF statistics for the short test with no ability difference. The figure contains horizontal three panels, one for each item discrimination parameter (a = .48, .60, or.75). Each panel has two vertical scatterplots: The upper scatterplot presents the bias and RMSD from the WSS MH DIF analysis with naive binning; the lower scatterplot represents the bias and RMSD from the SSS MH DIF analysis. *Note*. DIF = differential item functioning, MH = mantel–Haenszel, RMSD = root mean square difference, SSS = simple sum scores, WSS = weighted sum scores.

a = .48, .60, or.75. Comparing Figure 3 with Figure 1 or Tables B1–B3 with Tables C1–C3 for either WSS or SSS MH DIF statistics, the magnitude of bias or RMSD is similar under the same sample size condition. There is no apparent deterioration of SSS MH DIF statistics due to a shorter test when there is no ability difference.

The Case of Large Ability Difference

The bias and RMSD for the MH DIF statistics in the case of the short test with large ability difference are presented in Figure 4. As in the case of the long test with a large ability difference, positive or negative biases are observed for SSS MH DIF statistics in the panel for a = .48 or.75. As the sample size increase, the magnitude of the biases decreases. A close comparison of Figures 2 and 4 or Tables B6 and C6 shows that even with the largest sample sizes, the magnitude of such positive or negative bias on the short test is larger than that on the long test. In contrast, the bias of WSS MH DIF statistics on the short test is still around zero.

Unlike previous cases where the item difficulty or item difficulty difference parameters show no impact on MH DIF statistics, here we observed minor impacts of the item difficulty difference parameter on WSS MH DIF statistics. They are shown in Figure 5 for the simulation conditions for the largest sample sizes (i.e., the focal group has 4,800 test takers); in each scatterplot, the items were further grouped by their item difficulty difference parameters (d = 0.25, d = 0, or d = −0.25) between the focal and reference groups. One can think of Figure 5 as a magnified picture of Figure 4 in which only the largest sample sizes conditions are presented. In Figure 5, the most noticeable effect is still the item discrimination parameter on SSS MH DIF statistics (the scatterplots in the lower panels); that is, the positive or negative bias is observed with a = .48 or .75. The bias has no relationship with the item difficulty difference (d) between the focal and reference groups. Items with a = .48 have a positive bias for SSS MH DIF statistics, and items with a = .75, a negative bias .

However, a close look at the scatterplots in the top panels of Figure 5 shows that items with d = .25 have a slightly positive bias, items with d = −.25 have a slightly negative bias, and only the no-DIF items (the ones in the middle) have a close to zero bias for WSS MH DIF statistics. These minor biases are less evident with the long test for WSS MH DIF statistics. Thus, with large ability difference, the primary source of variation of SSS MH DIF statistics from DMI is still the item discrimination parameter. However, when the test length is short, the item difficulty difference parameter may have some impact on WSS MH DIF statistics, but the impact on WSS is much less than on SSS MH DIF statistics.

Summary and Discussion

Guo and Dorans (2019, 2020) generalized Holland and Thayer's (1988) theoretical DIF results from Rasch model to the 2PL model and showed that using weighted sum scores as the matching variable can close the gap between the

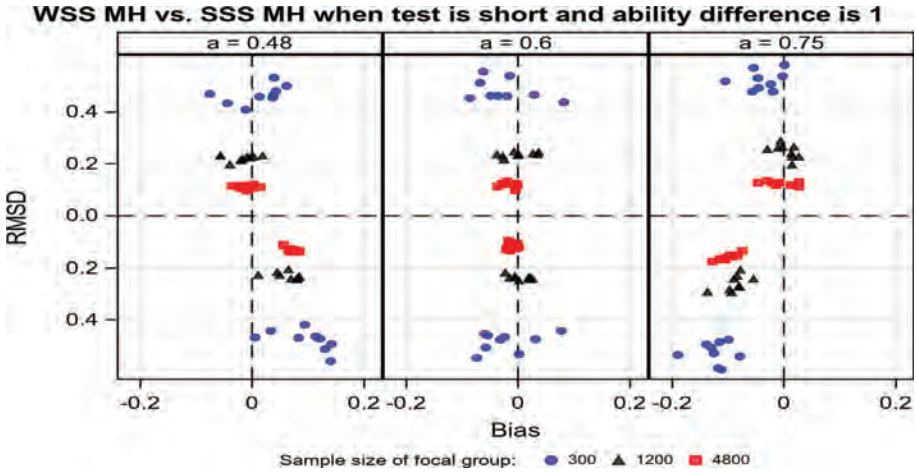

Figure 4 Bias and RMSD of MH DIF statistics for the short test with large ability difference. The figure contains horizontal three panels, one for each item discrimination parameter (a = .48, .60, or .75). Each panel has two vertical scatterplots: The upper scatterplot presents the bias and RMSD from the WSS MH DIF analysis with naive binning; the lower scatterplot represents the bias and RMSD from the SSS MH DIF analysis. *Note*. DIF = differential item functioning, MH = Mantel–Haenszel, RMSD = root mean square difference, SSS = simple sum scores, WSS = weighted sum scores.

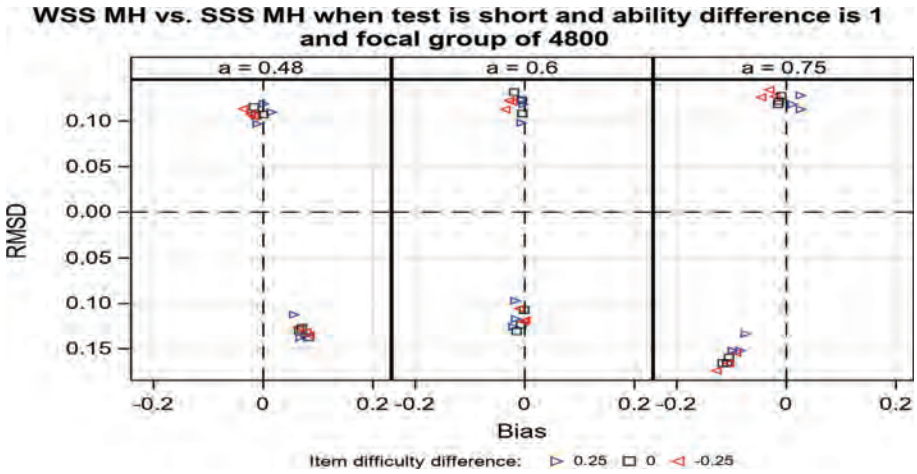

Figure 5 Bias and RMSD of MH DIF statistics for the short test with large ability difference and large sample sizes. The figure contains three horizontal panels, one for each item discrimination parameter (a = .48, .60, or .75). Each panel has two vertical scatterplots: The upper scatterplot presents the bias and RMSD from the WSS MH DIF analysis with naive binning; the lower scatterplot represents the bias and RMSD from the SSS MH DIF analysis. *Note*. DIF = differential item functioning, MH = Mantel–Haenszel, RMSD = root mean square difference, SSS = simple sum scores, WSS = weighted sum scores.

observed-score based DIF and latent-ability based DMI under some circumstances. Using an analytic method under the assumption that the 2PL model held for item responses, they derived population results for both the DIF based on observed scores (i.e., MH DIF statistics) and DMI based on latent ability. To overcome the computational intensity associated with the large number of weighted sum scores, they further proposed three binning methods to make the weighted DIF statistics practically feasible. With their analytic results and a real data example, they pointed out four factors that could affect the differences between DIF (or the MH DIF statistic) and DMI. The four factors are (a) ability differences between the focal group and reference group, (b) item discrimination parameters, (c) item difficulty parameters, and (d) test length.

To explore the operational use of the weighted DIF statistics, we used the same test design and treated the population DMI values as our criteria in this simulation study. We further manipulated the sample size condition, which is a critical factor in operational use that could affect both the item discrimination parameter estimation and operational MH DIF statistics. Results in this study were consistent with those in Guo and Dorans (2019, 2020) in terms of the impact of test

length, ability difference, item discrimination, and difficulty parameters on MH DIF analysis when the item responses follow the 2PL model.

The test length mattered for both WSS and SSS MH DIF statistics. The longer the test, the smaller the bias and RMSD. This finding can be explained by the difference in test reliability of long and short tests. In this study, when the ability difference was zero, the reliability was .84 for the short test and .91 for the long test in the total sample; when the ability difference was one, the reliability was .86 for the short test and .93 for the long test in the total sample. Tests with larger reliability provide a better approximation of the true proficiency via either simple sum scoring or weighted sum scoring, which lead to better agreement between DIF and DMI. However, the accelerated convergence rate of WSS MH DIF statistics to DMI in longer tests was not observed in simulations, compared to that in SSS MH DIF statistics. That is, WSS MH DIF statistics approximates DMI better when the test length is longer, at a similar rate to SSS MH DIF statistics. This may have resulted from the fact that more item parameters needed to be estimated for the long test with the same sample size, and thus larger item parameter estimation errors were present in WSS MH DIF statistics for the longer test.

Ability differences affected the difference between DIF and DMI the most. When there is no ability difference, both WSS MH and SSS MH DIF statistics were close to each other and close to the DMI measure. When the ability differences were large, we observed that the bias of SSS MH DIF statistics was not zero even for items that had no DIF. In contrast, the bias of WSS MH DIF statistics for all items was around zero. Furthermore, when the focal group was less able than the reference group, the bias of SSS MH DIF statistics depended on the size of the item discrimination parameter for the item with DMI. The bias was positive when the item's discrimination parameter was smaller than the average discrimination of the test, and the bias was negative when the item's discrimination parameter was larger than the average of the test. In contrast, WSS MH DIF statistics showed no such bias. Other item properties, such as item difficulty differences between the focal and reference groups had little impact on MH DIF statistics when the test was long. However, when the test was short and ability difference was large, WSS MH DIF statistics were slightly impacted by the item difficulty differences between the focal and reference groups for the studied items, but that impact was much larger for SSS MH DIF statistics.

With regard to the impact of sample sizes, the results of this study are consistent with previous DIF studies in that the larger the sample size was, the smaller the RMSD of SSS MH DIF statistics from DMI. This study finds that this finding also holds true for WSS MH DIF statistics. Note that the impact of sample sizes to WSS MH DIF statistics are twofold: one impact is on the estimated item discrimination parameters, which are critical elements in WSS MH DIF statistics. If the sample size is too small and the parameters are poorly estimated, there may be no advantage to using WSS — there may even be a disadvantage. As to the sample size requirements to get accurate item parameter estimates of the 2PL model, the results of previous studies (e.g., Hulin et al., 1982; Şahin & Anıl, 2017; Stone, 1992) generally apply here. For example, for a test with at least 20 items, a minimum sample size of 500 is required. With adequate item discrimination parameter estimates, we also found that the sample size factor interacted with other factors, such as the ability difference. In case of a large ability difference, the larger the sample size, the smaller the RMSD between MH DIF statistics and DMI; however, it could not eliminate the bias observed with SSS MH DIF statistics when the item discrimination parameters are above or below the average item discrimination parameter in the test. In contrast, WSS MH DIF statistics have no such bias.

The binning method has the least impact on WSS MH DIF statistics. With the same simulation condition, the three binning methods produced about the same results. Considering the easy implementation of the naive binning method (no need to equate WSS to SSS), we recommend the use of naive binning in operational settings for WSS MH DIF statistics.

From a practical perspective, our simulation results provided some support for the use of traditional MH DIF (i.e., SSS) statistics for tests that are reasonably reliable and have sufficiently large sample sizes, mainly because the bias is very small on the MH DIF scale (Zwick, 2012). On the other hand, our results also showed that in operational settings, if the 2PL model is used for score linking or score reporting, it would be more consistent to use the weighted DIF statistics to flag DIF items, especially when the ability difference is large, the test length is short, and the test has large variation in item discrimination parameters.

Although this study evaluated the robustness of weighted DIF statistics performance under various situations and explored the feasibility of the weighted DIF measure in an operational-like setting, a number of issues need further research. For example, this paper considered the use of MH as the DIF measure. The weighted sum scores could also be applied with other DIF methods (e.g., the logistic regression procedure, Swaminathan & Rogers, 1990). In addition, if the testing program uses an IRT model to produce reliable latent ability estimates, a binning method can be applied to latent ability estimates to compute the weighted DIF statistics that matches on the latent ability estimates. Comparison

between the weighted DIF statistics matching on weighted sum scores and latent ability estimates may merit investigation as well, in terms of accuracy and practical efficiency in operation. Moreover, even though the test design used in the study mimicked a real operational testing program, we ignored other factors that could affect the use of the weighted DIF statistics, such as model-data fit. When data do not fit a 2PL model, because of the estimation error associated with the item discrimination parameters, the simple sum score method might work better.

References

Bock, R. D., & Zimowski, M. F. (1997). Multiple group IRT. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 433 – 448). Springer. https://doi.org/10.1007/978-1-4757-2691-6_25

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Sage.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially funcioning test items. *Educational Measurement: Issues and Practice*, *17*(1), 31 – 44. https://doi.org/10.1111/j.1745-3992.1998.tb00619.x

Donoghue, J., Holland, P., & Thayer, D. (1993). A Monte Carlo study of factors that affect the Mantel-Haenszel and standardization measures of differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 137 – 166). Erlbaum.

Dorans, N. J., & Kulick, 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*(4), 355 – 368. https://doi.org/10.1111/j.1745-3984.1986.tb00255.x

Guo, H., & Dorans, N. J. (2019). *Observed scores as matching variables in differential item functioning under the one- and two-parameter logistic models: Population results* (Research Report No. RR-19-06). ETS. https://doi.org/10.1002/ets2.12243

Guo, H., & Dorans, N. J. (2020). Using the weighted sum scores to close the gap between DIF practice and theory. *Journal of Educational Measurement*, *57*(4), 484 – 510. https://doi.org/10.1111/jedm.12258

Harris, D. (1989). Comparion of 1-, 2-, and 3-parameter IRT models. *Educational Measurement: Issues and Practice*, *8*(1), 35 – 41. https://doi.org/10.1111/j.1745-3992.1989.tb00313.x

Harwell, M. R., & Janosky, J. E. (1991). An empirical study of the effects of small datasets and varying prior variances on item parameter estimation in BILOG. *Applied Psychological Measurement*, *15*(3), 279 – 291. https://doi.org/10.1177/014662169101500308

Holland, P., & Thayer, D. (1988). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 129 – 145). Erlbaum.

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two- and three-parameter logistic item characteristic curves: A Monte Carlo study. *Applied Psychological Measurement*, *6*(3), 249 – 260. https://doi.org/10.1177/014662168200600301

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices*. Springer.

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

Roussos, L. A., Schnipke, D. L., & Pashley, P. J. (1999). A generalized formula for the Mantel-Haenszel differential item functioning parameter. *Journal of Educational and Behaviorial Statistics*, *24*(3), 293 – 322. https://doi.org/10.3102/10769986024003293

Şahin, A., & Anıl, D. (2017). The effects of test length and sample size on item parameters in item response theory. *Educational Sciences: Theory & Practice*, *17*(1), 321 – 335. https://doi.org/10.12738/estp.2017.1.0270

Stone, C. A. (1992). Recovery of marginal maximum likelihood estimates in the two-parameter logistic reponse model: An evaluation of MULTILOG. *Appled Psychological Measurement*, *16*(1), 1 – 16. https://doi.org/10.1177/014662169201600101

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*(4), 361 – 370. https://doi.org/10.1111/j.1745-3984.1990.tb00754.x

Thissen, D., Steinberg, L., & Wainer, H. (1988). Detection of differential item functioning using the parameters of item response theory models. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 67 – 113). Erlbaum.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 147 – 169). Erlbaum.

Zwick, R. (2012). *A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement* (Research Report No. RR-12-08). ETS. https://doi.org/10.1002/j.2333-8504.2012.tb02290.x

Zwick, R., Thayer, D. T., & Lewis, C. (2000). Using loss functions for DIF detection: An empirical Bayes approach. *Journal of Educational and Behavioral Statistics*, *25*(2), 225 – 247. https://doi.org/10.3102/10769986025002225

Zwick, R., Thayer, D. T., & Mazzeo, J. (1997). Descriptive and inferential procedures for assessing differential item functioning in polytomous items. *Applied Measurement in Education*, *10*(4), 321 – 334. https://doi.org/10.1207/s15324818ame1004_2

Appendix A

Item Discrimination Parameter Estimates

For each item in Table 1, the bias and RMSD of the item discrimination parameter estimates are presented in Tables A1 – A3. Each table represents a sample size condition: When the focal group sample size was 300, 1,200, and 4,800, respectively. In each table, the bias and RMSD of four conditions are reported: long test length with no ability difference (Long_ABD0), long test length with large ability difference (Long_ABD1); short test length with no ability difference (Short_ABD0); and short test length with large ability difference (Short_ABD1).

When we conducted the item calibrations, a single population IRT model was used when the focal and reference group test takers were sampled from two different distributions, rather than a multiple-group IRT model (Bock & Zimowski, 1997). Biases are shown in Tables A1 – A3 for the simulation conditions when there are large ability difference, including both the long and short tests, and each of the sample sizes conditions. In contrast, when there is no ability difference between the two groups, the magnitude of RMSD is generally smaller and the biases of the item discrimination parameter estimates are closer to zero.

Consistent with previous findings on the impact of sample size and test length on item discrimination parameter estimates with the 2PL model (e.g., Harwell & Janosky, 1991; Hulin et al., 1982), smaller sample sizes and shorter test length yielded less accurate item discrimination parameter estimate when there is no ability difference between the focal and reference groups. With the largest sample size in this study (i.e., 3,600 test takers in the focal group and 4,800 test takers in the reference group), both the bias and RMSD of the item discrimination parameter estimates were close to zero.

Table A1 Bias and RMSD of Item Discrimination Parameter Estimate for Focal Group Sample Size of 300

				Long_ABD0		Long_ABD1		Short_ABD0		Short_ABD1	
Item	*a*	*b*	*d*	Bias	RMSD	Bias	RMSD	Bias	RMSD	Bias	RMSD
1	0.48	−1	−0.25	0.010	0.060	0.066	0.090	0.013	0.065	0.071	0.096
2	0.48	−1	0	0.011	0.059	0.067	0.092	0.014	0.062	0.070	0.094
3	0.48	−1	0.25	0.009	0.059	0.066	0.090	0.015	0.065	0.075	0.100
4	0.48	0	−0.25	0.008	0.056	0.066	0.088	0.014	0.061	0.069	0.093
5	0.48	0	0	0.009	0.057	0.065	0.087	0.014	0.060	0.072	0.095
6	0.48	0	0.25	0.010	0.059	0.065	0.087	0.016	0.061	0.073	0.096
7	0.48	1	−0.25	0.009	0.058	0.064	0.088	0.013	0.064	0.070	0.095
8	0.48	1	0	0.010	0.059	0.064	0.088	0.014	0.064	0.071	0.095
9	0.48	1	0.25	0.009	0.059	0.065	0.089	0.011	0.062	0.072	0.096
10	0.6	−1	−0.25	0.005	0.067	0.074	0.102	0.010	0.071	0.083	0.112
11	0.6	−1	0	0.004	0.067	0.075	0.103	0.009	0.073	0.083	0.111
12	0.6	−1	0.25	0.004	0.067	0.076	0.104	0.009	0.070	0.086	0.115
13	0.6	0	−0.25	0.003	0.063	0.075	0.100	0.009	0.067	0.082	0.108
14	0.6	0	0	0.004	0.063	0.074	0.099	0.010	0.067	0.081	0.107
15	0.6	0	0.25	0.004	0.063	0.075	0.101	0.009	0.069	0.085	0.111
16	0.6	1	−0.25	0.003	0.066	0.073	0.101	0.009	0.071	0.080	0.109
17	0.6	1	0	0.002	0.066	0.074	0.101	0.010	0.071	0.082	0.110
18	0.6	1	0.25	0.006	0.067	0.074	0.102	0.009	0.072	0.082	0.111
19	0.75	−1	−0.25	−0.002	0.078	0.087	0.121	0.007	0.083	0.097	0.131
20	0.75	−1	0	−0.001	0.079	0.089	0.122	0.007	0.082	0.098	0.132
21	0.75	−1	0.25	−0.001	0.077	0.089	0.120	0.007	0.084	0.099	0.133
22	0.75	0	−0.25	−0.001	0.072	0.087	0.116	0.004	0.077	0.098	0.128
23	0.75	0	0	0.001	0.073	0.088	0.116	0.003	0.077	0.098	0.128
24	0.75	0	0.25	−0.002	0.072	0.088	0.117	0.005	0.077	0.099	0.127
25	0.75	1	−0.25	−0.002	0.079	0.084	0.116	0.004	0.084	0.093	0.127
26	0.75	1	0	−0.004	0.077	0.084	0.117	0.002	0.081	0.095	0.129
27	0.75	1	0.25	−0.005	0.078	0.087	0.119	0.003	0.083	0.096	0.128

Note. Bias and RMSD were calculated across 100 replications. RMSD = root mean square difference; Long_ABD0 = long test length with no ability difference; Long_ABD1 = long test length with large ability difference; Short_ABD0 = short test length with no ability difference; Short_ABD1 = short test length with large ability difference.

Table A2 Bias and RMSD of Item Discrimination Parameter Estimate for Focal Group Sample Size of 1,200

				Long_ABD0		Long_ABD1		Short_ABD0		Short_ABD1	
Item	*a*	*b*	*d*	Bias	RMSD	Bias	RMSD	Bias	RMSD	Bias	RMSD
1	0.48	−1	−0.25	0.003	0.030	0.059	0.067	0.006	0.033	0.065	0.073
2	0.48	−1	0	0.002	0.031	0.059	0.067	0.008	0.034	0.065	0.073
3	0.48	−1	0.25	0.002	0.030	0.059	0.067	0.009	0.034	0.068	0.075
4	0.48	0	−0.25	0.002	0.029	0.058	0.065	0.007	0.032	0.064	0.071
5	0.48	0	0	0.002	0.030	0.059	0.066	0.008	0.032	0.065	0.072
6	0.48	0	0.25	0.002	0.029	0.059	0.066	0.006	0.032	0.066	0.074
7	0.48	1	−0.25	0.002	0.031	0.057	0.065	0.007	0.034	0.064	0.072
8	0.48	1	0	0.002	0.031	0.057	0.065	0.006	0.032	0.064	0.072
9	0.48	1	0.25	0.002	0.030	0.059	0.067	0.007	0.034	0.064	0.072
10	0.6	−1	−0.25	0.001	0.034	0.070	0.079	0.008	0.037	0.079	0.087
11	0.6	−1	0	0.002	0.035	0.072	0.081	0.008	0.037	0.081	0.090
12	0.6	−1	0.25	0.001	0.034	0.073	0.081	0.007	0.037	0.082	0.091
13	0.6	0	−0.25	0.001	0.032	0.071	0.079	0.006	0.035	0.079	0.087
14	0.6	0	0	0.001	0.032	0.072	0.080	0.008	0.036	0.080	0.087
15	0.6	0	0.25	0.001	0.033	0.072	0.080	0.007	0.034	0.081	0.088
16	0.6	1	−0.25	0.000	0.034	0.069	0.078	0.005	0.037	0.077	0.085
17	0.6	1	0	0.000	0.034	0.069	0.078	0.007	0.038	0.077	0.086
18	0.6	1	0.25	0.001	0.034	0.070	0.079	0.006	0.036	0.079	0.087
19	0.75	−1	−0.25	0.000	0.040	0.088	0.097	0.007	0.043	0.098	0.109
20	0.75	−1	0	0.000	0.040	0.090	0.099	0.007	0.044	0.099	0.109
21	0.75	−1	0.25	0.000	0.039	0.090	0.099	0.009	0.043	0.102	0.111
22	0.75	0	−0.25	0.000	0.036	0.088	0.096	0.006	0.040	0.098	0.107
23	0.75	0	0	0.000	0.037	0.089	0.097	0.007	0.040	0.099	0.107
24	0.75	0	0.25	0.000	0.037	0.089	0.097	0.006	0.040	0.101	0.109
25	0.75	1	−0.25	−0.002	0.040	0.085	0.095	0.006	0.043	0.095	0.105
26	0.75	1	0	−0.001	0.040	0.086	0.095	0.005	0.043	0.095	0.105
27	0.75	1	0.25	−0.002	0.039	0.086	0.095	0.006	0.042	0.097	0.107

Note. Bias and RMSD were calculated across 100 replications. RMSD = root mean square difference; Long_ABD0 = long test length with no ability difference; Long_ABD1 = long test length with large ability difference; Short_ABD0 = short test length with no ability difference; Short_ABD1 = short test length with large ability difference.

Table A3 Bias and RMSD of Item Discrimination Parameter Estimate for Focal Group Sample Size of 4,800

				Long_ABD0		Long_ABD1		Short_ABD0		Short_ABD1	
Item	*a*	*b*	*d*	Bias	RMSD	Bias	RMSD	Bias	RMSD	Bias	RMSD
1	0.48	−1	−0.25	0.001	0.015	0.056	0.059	0.006	0.017	0.063	0.065
2	0.48	−1	0	0.001	0.015	0.057	0.059	0.006	0.018	0.063	0.065
3	0.48	−1	0.25	0.001	0.016	0.058	0.060	0.006	0.017	0.065	0.067
4	0.48	0	−0.25	0.001	0.015	0.055	0.057	0.006	0.017	0.062	0.064
5	0.48	0	0	0.001	0.015	0.056	0.058	0.005	0.016	0.063	0.065
6	0.48	0	0.25	0.001	0.014	0.057	0.059	0.006	0.017	0.064	0.066
7	0.48	1	−0.25	0.000	0.015	0.055	0.057	0.006	0.017	0.061	0.063
8	0.48	1	0	0.001	0.015	0.056	0.058	0.005	0.017	0.062	0.064
9	0.48	1	0.25	0.000	0.015	0.056	0.058	0.005	0.017	0.063	0.065
10	0.6	−1	−0.25	0.001	0.017	0.071	0.073	0.006	0.019	0.079	0.082
11	0.6	−1	0	0.001	0.017	0.071	0.074	0.006	0.020	0.079	0.082
12	0.6	−1	0.25	0.001	0.017	0.072	0.074	0.007	0.019	0.081	0.083
13	0.6	0	−0.25	0.001	0.016	0.070	0.072	0.007	0.018	0.078	0.080
14	0.6	0	0	0.000	0.016	0.070	0.072	0.007	0.019	0.079	0.082
15	0.6	0	0.25	0.001	0.016	0.071	0.073	0.006	0.019	0.081	0.083
16	0.6	1	−0.25	0.000	0.017	0.068	0.070	0.006	0.020	0.076	0.079
17	0.6	1	0	0.000	0.017	0.069	0.071	0.006	0.019	0.077	0.079
18	0.6	1	0.25	0.000	0.017	0.070	0.072	0.006	0.019	0.078	0.080
19	0.75	−1	−0.25	0.001	0.020	0.088	0.091	0.009	0.023	0.099	0.101
20	0.75	−1	0	0.001	0.020	0.089	0.091	0.009	0.023	0.100	0.102
21	0.75	−1	0.25	0.001	0.020	0.090	0.092	0.008	0.023	0.102	0.105
22	0.75	0	−0.25	0.000	0.018	0.087	0.089	0.008	0.021	0.098	0.100
23	0.75	0	0	0.000	0.019	0.088	0.090	0.008	0.021	0.099	0.101
24	0.75	0	0.25	0.000	0.018	0.089	0.091	0.008	0.022	0.101	0.103
25	0.75	1	−0.25	−0.001	0.020	0.085	0.087	0.006	0.022	0.095	0.097
26	0.75	1	0	−0.001	0.020	0.086	0.088	0.007	0.022	0.096	0.099
27	0.75	1	0.25	0.000	0.019	0.086	0.089	0.006	0.022	0.097	0.100

Note. Bias and RMSD were calculated across 100 replications. RMSD = root mean square difference; Long_ABD0 = long test length with no ability difference; Long_ABD1 = long test length with large ability difference; Short_ABD0 = short test length with no ability difference; Short_ABD1 = short test length with large ability difference.

Appendix B

MH DIF Statistics for the Long Test

Table B1 Bias and RMSD of MH DIF Statistics for Long Test With No Ability Difference for Focal Group Sample Size of 300

Test version				Simple sum score		Weighted sum score					
						Naive binning		Linear binning		Equipercentile binning	
Item	*a*	*b*	*d*	Bias	RMSD	Bias	RMSD	Bias	RMSD	Bias	RMSD
1	0.48	−1	0.25	−0.046	0.415	−0.041	0.430	−0.045	0.435	−0.050	0.437
2	0.48	−1	0	0.058	0.405	0.086	0.408	0.083	0.422	0.084	0.412
3	0.48	−1	−0.25	−0.047	0.478	−0.054	0.489	−0.061	0.492	−0.072	0.476
4	0.48	0	0.25	0.037	0.423	0.053	0.395	0.041	0.392	0.049	0.389
5	0.48	0	0	−0.004	0.420	−0.011	0.432	−0.002	0.420	0.003	0.429
6	0.48	0	−0.25	0.016	0.409	0.025	0.405	0.014	0.399	0.022	0.399
7	0.48	1	0.25	0.015	0.370	0.010	0.373	0.013	0.362	0.008	0.376
8	0.48	1	0	−0.017	0.421	−0.009	0.436	−0.007	0.434	−0.015	0.425
9	0.48	1	−0.25	−0.027	0.408	−0.023	0.431	−0.027	0.420	−0.030	0.417
10	0.6	−1	0.25	−0.038	0.457	−0.053	0.451	−0.052	0.443	−0.048	0.452
11	0.6	−1	0	0.061	0.406	0.061	0.403	0.050	0.415	0.055	0.416
12	0.6	−1	−0.25	0.098	0.489	0.107	0.474	0.103	0.489	0.103	0.475
13	0.6	0	0.25	0.004	0.422	−0.002	0.418	0.012	0.414	0.008	0.422
14	0.6	0	0	−0.063	0.428	−0.041	0.441	−0.056	0.440	−0.056	0.443
15	0.6	0	−0.25	0.050	0.425	0.055	0.411	0.061	0.424	0.061	0.428
16	0.6	1	0.25	0.018	0.449	−0.001	0.455	−0.007	0.472	−0.002	0.457
17	0.6	1	0	−0.050	0.464	−0.065	0.428	−0.051	0.440	−0.062	0.441
18	0.6	1	−0.25	−0.061	0.491	−0.071	0.497	−0.069	0.493	−0.062	0.494
19	0.75	−1	0.25	0.003	0.461	0.008	0.466	0.024	0.471	0.015	0.476
20	0.75	−1	0	−0.060	0.447	−0.068	0.468	−0.066	0.450	−0.070	0.461
21	0.75	−1	−0.25	0.024	0.446	0.024	0.458	0.016	0.449	0.035	0.448
22	0.75	0	0.25	−0.012	0.429	−0.002	0.430	−0.007	0.421	−0.004	0.418
23	0.75	0	0	−0.043	0.472	−0.031	0.484	−0.036	0.497	−0.040	0.498
24	0.75	0	−0.25	0.084	0.430	0.080	0.440	0.088	0.433	0.086	0.450
25	0.75	1	0.25	0.000	0.529	−0.003	0.527	−0.017	0.524	−0.019	0.526
26	0.75	1	0	−0.063	0.480	−0.047	0.494	−0.054	0.477	−0.052	0.462
27	0.75	1	−0.25	0.029	0.483	0.037	0.505	0.026	0.520	0.031	0.520

Note. MH DIF statistics were calculated across 100 replications. MH = Mantel–Haenszel; RMSD = root mean square difference.

Table B2 Bias and RMSD of MH DIF Statistics for Long Test With No Ability Difference for Focal Group Sample Size of 1,200

Test version				Simple sum score		Weighted sum score					
						Naive binning		Linear binning		Equipercentile binning	
Item	*a*	*b*	*d*	Bias	RMSD	Bias	RMSD	Bias	RMSD	Bias	RMSD
1	0.48	−1	0.25	0.021	0.179	0.022	0.176	0.022	0.174	0.022	0.174
2	0.48	−1	0	0.022	0.209	0.023	0.209	0.025	0.209	0.024	0.209
3	0.48	−1	−0.25	−0.057	0.213	−0.059	0.213	−0.061	0.214	−0.057	0.213
4	0.48	0	0.25	−0.001	0.181	0.004	0.183	0.001	0.187	0.002	0.184
5	0.48	0	0	−0.027	0.187	−0.026	0.189	−0.028	0.189	−0.025	0.189
6	0.48	0	−0.25	0.010	0.199	0.012	0.197	0.011	0.201	0.011	0.200
7	0.48	1	0.25	−0.015	0.193	−0.016	0.197	−0.015	0.191	−0.017	0.193
8	0.48	1	0	−0.001	0.203	−0.003	0.202	−0.003	0.202	−0.004	0.201
9	0.48	1	−0.25	−0.001	0.200	−0.001	0.198	0.002	0.197	0.000	0.198
10	0.6	−1	0.25	0.005	0.237	0.002	0.233	0.004	0.234	0.006	0.231
11	0.6	−1	0	0.030	0.232	0.030	0.229	0.028	0.230	0.028	0.231
12	0.6	−1	−0.25	−0.014	0.231	−0.012	0.233	−0.016	0.230	−0.016	0.233
13	0.6	0	0.25	0.025	0.181	0.023	0.178	0.023	0.179	0.022	0.179
14	0.6	0	0	0.007	0.195	0.009	0.196	0.007	0.199	0.007	0.198
15	0.6	0	−0.25	−0.007	0.196	−0.003	0.194	−0.004	0.196	−0.006	0.197
16	0.6	1	0.25	−0.008	0.234	−0.006	0.230	−0.006	0.230	−0.004	0.229
17	0.6	1	0	0.018	0.213	0.020	0.211	0.020	0.210	0.017	0.210
18	0.6	1	−0.25	−0.003	0.194	−0.006	0.191	−0.006	0.194	−0.004	0.196
19	0.75	−1	0.25	−0.012	0.220	−0.021	0.223	−0.022	0.225	−0.026	0.228
20	0.75	−1	0	−0.018	0.209	−0.018	0.213	−0.015	0.211	−0.018	0.213
21	0.75	−1	−0.25	0.015	0.232	0.017	0.235	0.019	0.235	0.015	0.236
22	0.75	0	0.25	0.028	0.207	0.029	0.204	0.027	0.208	0.029	0.208
23	0.75	0	0	0.003	0.184	0.002	0.182	0.002	0.181	0.003	0.181
24	0.75	0	−0.25	−0.078	0.239	−0.073	0.239	−0.073	0.238	−0.071	0.239
25	0.75	1	0.25	−0.046	0.255	−0.055	0.255	−0.053	0.258	−0.055	0.255
26	0.75	1	0	−0.016	0.234	−0.013	0.236	−0.013	0.236	−0.013	0.233
27	0.75	1	−0.25	0.003	0.224	0.007	0.227	0.006	0.226	0.003	0.225

Note. MH DIF statistics were calculated across 100 replications. MH = Mantel–Haenszel; RMSD = root mean square difference.

Table B3 Bias and RMSD of MH DIF Statistics for Long Test With No Ability Difference for Focal Group Sample Size of 4,800

Test version				Simple sum score		Weighted sum score					
						Naive binning		Linear binning		Equipercentile binning	
Item	*a*	*b*	*d*	Bias	RMSD	Bias	RMSD	Bias	RMSD	Bias	RMSD
1	0.48	−1	0.25	−0.006	0.110	−0.005	0.111	−0.004	0.110	−0.004	0.110
2	0.48	−1	0	−0.004	0.116	−0.002	0.114	−0.001	0.114	−0.002	0.113
3	0.48	−1	−0.25	0.009	0.105	0.008	0.106	0.008	0.105	0.009	0.105
4	0.48	0	0.25	−0.002	0.097	−0.001	0.097	0.000	0.098	0.000	0.097
5	0.48	0	0	0.006	0.092	0.005	0.091	0.005	0.090	0.005	0.091
6	0.48	0	−0.25	−0.009	0.106	−0.011	0.106	−0.011	0.106	−0.011	0.106
7	0.48	1	0.25	0.007	0.107	0.009	0.107	0.009	0.107	0.008	0.107
8	0.48	1	0	0.007	0.098	0.007	0.098	0.007	0.098	0.007	0.098
9	0.48	1	−0.25	−0.007	0.110	−0.007	0.109	−0.008	0.109	−0.008	0.109
10	0.6	−1	0.25	−0.006	0.119	−0.006	0.120	−0.006	0.120	−0.006	0.120
11	0.6	−1	0	0.012	0.109	0.013	0.108	0.013	0.108	0.014	0.108
12	0.6	−1	−0.25	0.017	0.127	0.015	0.127	0.015	0.127	0.015	0.127
13	0.6	0	0.25	−0.008	0.113	−0.008	0.113	−0.008	0.113	−0.008	0.112
14	0.6	0	0	0.009	0.078	0.009	0.078	0.009	0.079	0.009	0.080
15	0.6	0	−0.25	−0.001	0.094	−0.002	0.093	−0.001	0.094	−0.001	0.094
16	0.6	1	0.25	0.018	0.111	0.019	0.111	0.019	0.111	0.020	0.111
17	0.6	1	0	0.001	0.109	0.001	0.109	0.002	0.109	0.002	0.110
18	0.6	1	−0.25	0.002	0.102	0.003	0.102	0.002	0.103	0.002	0.103
19	0.75	−1	0.25	0.007	0.103	0.003	0.103	0.003	0.103	0.003	0.104
20	0.75	−1	0	0.014	0.113	0.014	0.114	0.013	0.113	0.014	0.113
21	0.75	−1	−0.25	0.003	0.119	0.007	0.119	0.006	0.120	0.006	0.119
22	0.75	0	0.25	−0.001	0.102	−0.006	0.104	−0.006	0.104	−0.006	0.104
23	0.75	0	0	−0.011	0.100	−0.010	0.100	−0.011	0.100	−0.010	0.100
24	0.75	0	−0.25	0.004	0.109	0.010	0.109	0.009	0.110	0.009	0.109
25	0.75	1	0.25	0.003	0.128	0.000	0.130	0.001	0.130	0.001	0.129
26	0.75	1	0	0.005	0.116	0.004	0.116	0.004	0.117	0.005	0.117
27	0.75	1	−0.25	0.009	0.108	0.012	0.108	0.013	0.108	0.012	0.107

Note. MH DIF statistics were calculated across 100 replications. MH = Mantel–Haenszel; RMSD = root mean square difference.

Table B4 Bias and RMSD of MH DIF Statistics for Long Test With Large Ability Difference for Focal Group Sample Size of 300

Test version				Simple sum score		Weighted sum score					
						Naive binning		Linear binning		Equipercentile binning	
Item	*a*	*b*	*d*	Bias	RMSD	Bias	RMSD	Bias	RMSD	Bias	RMSD
1	0.48	−1	0.25	0.108	0.449	0.065	0.462	0.072	0.452	0.058	0.452
2	0.48	−1	0	0.075	0.467	0.065	0.487	0.061	0.497	0.065	0.505
3	0.48	−1	−0.25	0.071	0.489	0.012	0.486	0.005	0.491	0.006	0.488
4	0.48	0	0.25	0.008	0.440	−0.019	0.433	−0.035	0.425	−0.030	0.430
5	0.48	0	0	0.007	0.396	−0.041	0.410	−0.031	0.403	−0.031	0.405
6	0.48	0	−0.25	0.055	0.446	−0.003	0.448	−0.002	0.440	−0.005	0.438
7	0.48	1	0.25	−0.015	0.505	−0.048	0.488	−0.035	0.487	−0.039	0.488
8	0.48	1	0	−0.040	0.483	−0.054	0.458	−0.065	0.482	−0.072	0.475
9	0.48	1	−0.25	0.015	0.445	−0.054	0.468	−0.049	0.466	−0.046	0.470
10	0.6	−1	0.25	0.070	0.455	0.063	0.461	0.067	0.463	0.058	0.462
11	0.6	−1	0	0.027	0.436	0.022	0.439	0.027	0.432	0.047	0.436
12	0.6	−1	−0.25	−0.029	0.530	−0.020	0.538	−0.020	0.533	−0.020	0.529
13	0.6	0	0.25	−0.010	0.492	−0.009	0.497	−0.003	0.483	−0.014	0.477
14	0.6	0	0	0.015	0.489	0.003	0.508	0.019	0.516	0.015	0.506
15	0.6	0	−0.25	0.016	0.455	−0.006	0.475	−0.001	0.463	−0.007	0.471
16	0.6	1	0.25	−0.038	0.571	−0.035	0.597	−0.034	0.571	−0.027	0.578
17	0.6	1	0	−0.029	0.532	−0.020	0.528	−0.006	0.525	−0.017	0.524
18	0.6	1	−0.25	0.078	0.474	0.067	0.481	0.048	0.463	0.052	0.463
19	0.75	−1	0.25	−0.058	0.569	−0.014	0.559	−0.018	0.558	−0.004	0.561
20	0.75	−1	0	0.069	0.488	0.130	0.504	0.137	0.496	0.121	0.485
21	0.75	−1	−0.25	−0.082	0.590	−0.044	0.595	−0.020	0.608	−0.012	0.604
22	0.75	0	0.25	−0.108	0.451	−0.042	0.445	−0.041	0.440	−0.047	0.445
23	0.75	0	0	−0.070	0.488	−0.023	0.457	−0.031	0.457	−0.037	0.463
24	0.75	0	−0.25	−0.068	0.433	−0.035	0.458	−0.035	0.458	−0.036	0.459
25	0.75	1	0.25	−0.154	0.534	−0.111	0.537	−0.112	0.517	−0.118	0.514
26	0.75	1	0	−0.042	0.572	0.023	0.559	0.013	0.570	0.017	0.570
27	0.75	1	−0.25	−0.089	0.535	−0.049	0.519	−0.064	0.512	−0.057	0.508

Note. MH DIF statistics were calculated across 100 replications. MH = Mantel–Haenszel; RMSD = root mean square difference.

Table B5 Bias and RMSD MH DIF Statistics for Long Test With Large Ability Difference for Focal Group Sample Size of 1,200

Test version				Simple sum score		Weighted sum score					
						Naive binning		Linear binning		Equipercentile binning	
Item	*a*	*b*	*d*	Bias	RMSD	Bias	RMSD	Bias	RMSD	Bias	RMSD
1	0.48	−1	0.25	0.010	0.215	−0.022	0.223	−0.026	0.223	−0.026	0.223
2	0.48	−1	0	0.002	0.226	−0.043	0.226	−0.040	0.228	−0.040	0.224
3	0.48	−1	−0.25	0.029	0.228	−0.025	0.224	−0.023	0.223	−0.024	0.224
4	0.48	0	0.25	0.035	0.225	−0.001	0.216	−0.003	0.217	−0.002	0.218
5	0.48	0	0	0.031	0.208	−0.016	0.199	−0.013	0.204	−0.013	0.204
6	0.48	0	−0.25	0.011	0.213	−0.044	0.218	−0.041	0.218	−0.043	0.219
7	0.48	1	0.25	0.046	0.257	0.008	0.254	0.011	0.255	0.011	0.255
8	0.48	1	0	0.031	0.244	−0.005	0.242	−0.006	0.241	−0.005	0.240
9	0.48	1	−0.25	0.050	0.230	0.004	0.220	0.002	0.218	0.002	0.218
10	0.6	−1	0.25	−0.021	0.232	−0.011	0.237	−0.012	0.232	−0.014	0.232
11	0.6	−1	0	0.016	0.249	0.021	0.250	0.015	0.247	0.016	0.249
12	0.6	−1	−0.25	−0.022	0.256	−0.030	0.255	−0.031	0.256	−0.028	0.259
13	0.6	0	0.25	−0.005	0.218	0.000	0.223	−0.002	0.219	0.001	0.219
14	0.6	0	0	−0.049	0.244	−0.056	0.249	−0.054	0.247	−0.055	0.250
15	0.6	0	−0.25	0.029	0.234	0.014	0.233	0.016	0.234	0.015	0.235
16	0.6	1	0.25	0.021	0.247	0.026	0.246	0.025	0.246	0.025	0.247
17	0.6	1	0	−0.006	0.263	−0.008	0.263	−0.007	0.260	−0.009	0.261
18	0.6	1	−0.25	0.025	0.244	0.013	0.243	0.013	0.247	0.013	0.247
19	0.75	−1	0.25	−0.059	0.245	0.007	0.240	0.000	0.240	−0.001	0.240
20	0.75	−1	0	−0.017	0.234	0.037	0.241	0.043	0.239	0.046	0.239
21	0.75	−1	−0.25	−0.092	0.267	−0.050	0.256	−0.052	0.256	−0.051	0.256
22	0.75	0	0.25	−0.056	0.235	−0.005	0.231	−0.003	0.229	−0.002	0.231
23	0.75	0	0	−0.070	0.261	−0.017	0.255	−0.022	0.254	−0.019	0.258
24	0.75	0	−0.25	−0.028	0.240	0.010	0.241	0.011	0.240	0.013	0.241
25	0.75	1	0.25	−0.029	0.269	0.032	0.273	0.035	0.265	0.034	0.267
26	0.75	1	0	−0.041	0.270	0.004	0.271	0.007	0.268	0.008	0.270
27	0.75	1	−0.25	−0.077	0.297	−0.042	0.294	−0.044	0.291	−0.044	0.288

Note. MH DIF statistics were calculated across 100 replications. MH = Mantel–Haenszel; RMSD = root mean square difference.

Table B6 Bias and RMSD MH DIF Statistics for Long Test With Large Ability Difference for Focal Group Sample Size of 4,800

Test version				Simple sum score		Weighted sum score					
						Naive binning		Linear binning		Equipercentile binning	
Item	*a*	*b*	*d*	Bias	RMSD	Bias	RMSD	Bias	RMSD	Bias	RMSD
1	0.48	−1	0.25	0.056	0.112	−0.010	0.097	−0.001	0.097	−0.001	0.097
2	0.48	−1	0	0.065	0.129	−0.018	0.116	−0.007	0.115	−0.007	0.115
3	0.48	−1	−0.25	0.065	0.126	−0.036	0.114	−0.024	0.111	−0.025	0.112
4	0.48	0	0.25	0.079	0.135	0.015	0.110	0.021	0.111	0.020	0.111
5	0.48	0	0	0.070	0.128	−0.013	0.107	−0.005	0.106	−0.005	0.106
6	0.48	0	−0.25	0.078	0.132	−0.023	0.108	−0.013	0.107	−0.012	0.107
7	0.48	1	0.25	0.068	0.137	0.001	0.119	0.009	0.119	0.009	0.119
8	0.48	1	0	0.083	0.137	0.000	0.107	0.010	0.108	0.009	0.108
9	0.48	1	−0.25	0.084	0.135	−0.018	0.105	−0.007	0.105	−0.008	0.104
10	0.6	−1	0.25	−0.021	0.126	−0.004	0.124	−0.006	0.124	−0.006	0.124
11	0.6	−1	0	−0.007	0.124	−0.005	0.123	−0.006	0.123	−0.006	0.123
12	0.6	−1	−0.25	0.001	0.119	−0.023	0.121	−0.019	0.121	−0.020	0.121
13	0.6	0	0.25	−0.018	0.097	−0.005	0.098	−0.006	0.098	−0.006	0.098
14	0.6	0	0	−0.001	0.107	−0.005	0.109	−0.005	0.109	−0.004	0.109
15	0.6	0	−0.25	−0.011	0.106	−0.036	0.113	−0.033	0.113	−0.033	0.113
16	0.6	1	0.25	−0.016	0.117	−0.002	0.119	−0.004	0.118	−0.004	0.118
17	0.6	1	0	−0.015	0.131	−0.020	0.132	−0.018	0.132	−0.018	0.131
18	0.6	1	−0.25	−0.002	0.120	−0.029	0.123	−0.026	0.122	−0.025	0.122
19	0.75	−1	0.25	−0.084	0.152	0.026	0.128	0.017	0.128	0.017	0.128
20	0.75	−1	0	−0.116	0.166	−0.015	0.119	−0.025	0.120	−0.026	0.121
21	0.75	−1	−0.25	−0.128	0.174	−0.046	0.127	−0.055	0.129	−0.055	0.129
22	0.75	0	0.25	−0.075	0.134	0.027	0.113	0.020	0.113	0.021	0.113
23	0.75	0	0	−0.105	0.160	−0.014	0.122	−0.021	0.122	−0.021	0.122
24	0.75	0	−0.25	−0.092	0.155	−0.020	0.127	−0.025	0.129	−0.024	0.128
25	0.75	1	0.25	−0.097	0.152	0.012	0.119	0.001	0.117	0.002	0.118
26	0.75	1	0	−0.107	0.166	−0.009	0.128	−0.019	0.128	−0.018	0.127
27	0.75	1	−0.25	−0.105	0.166	−0.031	0.135	−0.038	0.135	−0.039	0.135

Note. MH DIF statistics were calculated across 100 replications. MH = Mantel–Haenszel; RMSD = root mean square difference.

Appendix C

MH DIF Statistics for the Short Test

Table C1 Bias and RMSD of MH DIF Statistics for Short Test With No Ability Difference for Focal Group Sample Size of 300

Test version				Simple sum score		Weighted sum score					
						Naive binning		Linear binning		Equipercentile binning	
Item	*a*	*b*	*d*	Bias	RMSD	Bias	RMSD	Bias	RMSD	Bias	RMSD
1	0.48	−1	0.25	−0.026	0.421	−0.017	0.424	−0.022	0.426	−0.023	0.422
2	0.48	−1	0	0.036	0.437	0.046	0.447	0.040	0.433	0.046	0.427
3	0.48	−1	−0.25	0.056	0.422	0.046	0.401	0.057	0.417	0.057	0.414
4	0.48	0	0.25	0.052	0.401	0.060	0.395	0.061	0.400	0.061	0.400
5	0.48	0	0	−0.039	0.464	−0.033	0.457	−0.043	0.458	−0.040	0.461
6	0.48	0	−0.25	0.063	0.409	0.058	0.409	0.059	0.412	0.062	0.412
7	0.48	1	0.25	−0.006	0.375	−0.004	0.377	−0.007	0.372	−0.009	0.380
8	0.48	1	0	−0.012	0.418	−0.026	0.410	−0.017	0.405	−0.024	0.407
9	0.48	1	−0.25	0.036	0.409	0.028	0.404	0.036	0.412	0.039	0.405
10	0.6	−1	0.25	0.032	0.452	0.043	0.457	0.038	0.444	0.038	0.447
11	0.6	−1	0	−0.061	0.394	−0.060	0.394	−0.057	0.394	−0.060	0.399
12	0.6	−1	−0.25	0.090	0.428	0.103	0.410	0.097	0.424	0.098	0.422
13	0.6	0	0.25	0.054	0.391	0.041	0.389	0.046	0.382	0.049	0.385
14	0.6	0	0	0.014	0.452	0.010	0.452	0.009	0.443	0.006	0.445
15	0.6	0	−0.25	0.073	0.436	0.080	0.424	0.087	0.433	0.087	0.427
16	0.6	1	0.25	−0.017	0.445	−0.005	0.442	−0.017	0.449	−0.016	0.445
17	0.6	1	0	0.108	0.452	0.098	0.450	0.097	0.456	0.096	0.450
18	0.6	1	−0.25	0.010	0.368	0.022	0.379	0.015	0.369	0.011	0.377
19	0.75	−1	0.25	0.054	0.456	0.046	0.462	0.053	0.454	0.048	0.455
20	0.75	−1	0	−0.027	0.518	−0.026	0.536	−0.027	0.512	−0.027	0.512
21	0.75	−1	−0.25	−0.090	0.524	−0.078	0.544	−0.070	0.535	−0.075	0.533
22	0.75	0	0.25	0.020	0.376	0.019	0.393	0.017	0.399	0.013	0.396
23	0.75	0	0	0.123	0.444	0.126	0.449	0.128	0.448	0.120	0.444
24	0.75	0	−0.25	−0.042	0.422	−0.031	0.430	−0.033	0.424	−0.033	0.422
25	0.75	1	0.25	0.027	0.442	0.010	0.452	0.020	0.455	0.015	0.456
26	0.75	1	0	−0.027	0.427	−0.031	0.446	−0.028	0.440	−0.027	0.434
27	0.75	1	−0.25	−0.004	0.407	−0.001	0.414	−0.002	0.420	−0.004	0.418

Note. MH DIF statistics were calculated across 100 replications. MH = Mantel–Haenszel; RMSD = root mean square difference.

Table C2 Bias and RMSD of MH DIF Statistics for Short Test With No Ability Difference for Focal Group Sample Size of 1,200

Test version				Simple sum score		Weighted sum score					
						Naive binning		Linear binning		Equipercentile binning	
Item	*a*	*b*	*d*	Bias	RMSD	Bias	RMSD	Bias	RMSD	Bias	RMSD
1	0.48	−1	0.25	−0.030	0.208	−0.031	0.207	−0.029	0.207	−0.029	0.206
2	0.48	−1	0	−0.016	0.212	−0.017	0.211	−0.017	0.212	−0.017	0.210
3	0.48	−1	−0.25	−0.005	0.213	−0.009	0.213	−0.008	0.211	−0.009	0.210
4	0.48	0	0.25	0.031	0.192	0.034	0.193	0.034	0.191	0.034	0.191
5	0.48	0	0	0.017	0.212	0.018	0.215	0.017	0.215	0.017	0.215
6	0.48	0	−0.25	−0.026	0.192	−0.026	0.189	−0.029	0.189	−0.029	0.188
7	0.48	1	0.25	0.033	0.224	0.036	0.221	0.037	0.224	0.036	0.224
8	0.48	1	0	−0.011	0.226	−0.012	0.225	−0.011	0.222	−0.011	0.221
9	0.48	1	−0.25	−0.016	0.211	−0.014	0.205	−0.016	0.206	−0.016	0.206
10	0.6	−1	0.25	−0.053	0.211	−0.052	0.209	−0.054	0.210	−0.056	0.211
11	0.6	−1	0	−0.019	0.233	−0.018	0.235	−0.016	0.235	−0.016	0.235
12	0.6	−1	−0.25	−0.022	0.234	−0.024	0.235	−0.022	0.231	−0.022	0.231
13	0.6	0	0.25	−0.015	0.206	−0.013	0.209	−0.012	0.209	−0.014	0.209
14	0.6	0	0	−0.001	0.212	0.001	0.212	0.002	0.211	0.002	0.210
15	0.6	0	−0.25	−0.012	0.173	−0.011	0.173	−0.010	0.175	−0.011	0.174
16	0.6	1	0.25	−0.012	0.236	−0.010	0.236	−0.009	0.235	−0.008	0.234
17	0.6	1	0	−0.003	0.196	−0.007	0.196	−0.005	0.198	−0.005	0.199
18	0.6	1	−0.25	−0.019	0.229	−0.020	0.229	−0.017	0.226	−0.018	0.226
19	0.75	−1	0.25	0.043	0.215	0.033	0.220	0.037	0.221	0.037	0.220
20	0.75	−1	0	0.024	0.233	0.022	0.239	0.021	0.235	0.022	0.236
21	0.75	−1	−0.25	−0.001	0.222	0.001	0.224	0.000	0.222	0.000	0.222
22	0.75	0	0.25	−0.027	0.206	−0.032	0.209	−0.032	0.207	−0.032	0.208
23	0.75	0	0	−0.023	0.229	−0.023	0.229	−0.024	0.229	−0.025	0.230
24	0.75	0	−0.25	−0.016	0.209	−0.008	0.212	−0.010	0.210	−0.010	0.212
25	0.75	1	0.25	−0.025	0.229	−0.037	0.238	−0.035	0.233	−0.033	0.232
26	0.75	1	0	−0.035	0.239	−0.036	0.240	−0.038	0.241	−0.037	0.241
27	0.75	1	−0.25	−0.020	0.223	−0.014	0.227	−0.012	0.222	−0.012	0.222

Note. MH DIF statistics were calculated across 100 replications. MH = Mantel–Haenszel; RMSD = root mean square difference.

Table C3 Bias and RMSD of MH DIF Statistics for Short Test With No Ability Difference for Focal Group Sample Size of 4,800

Test version				Simple sum score		Weighted sum score					
						Naive binning		Linear binning		Equipercentile binning	
Item	*a*	*b*	*d*	Bias	RMSD	Bias	RMSD	Bias	RMSD	Bias	RMSD
1	0.48	−1	0.25	0.012	0.105	0.013	0.103	0.014	0.103	0.015	0.104
2	0.48	−1	0	−0.004	0.106	−0.005	0.105	−0.005	0.106	−0.005	0.106
3	0.48	−1	−0.25	−0.017	0.089	−0.021	0.090	−0.020	0.089	−0.020	0.089
4	0.48	0	0.25	−0.008	0.091	−0.005	0.089	−0.005	0.089	−0.005	0.089
5	0.48	0	0	−0.005	0.100	−0.004	0.099	−0.005	0.099	−0.005	0.098
6	0.48	0	−0.25	−0.003	0.086	−0.007	0.086	−0.006	0.086	−0.006	0.086
7	0.48	1	0.25	0.006	0.101	0.008	0.100	0.007	0.100	0.007	0.100
8	0.48	1	0	−0.001	0.108	−0.002	0.106	−0.001	0.107	−0.001	0.107
9	0.48	1	−0.25	0.012	0.105	0.009	0.103	0.009	0.104	0.009	0.103
10	0.6	−1	0.25	0.023	0.108	0.023	0.110	0.024	0.109	0.024	0.108
11	0.6	−1	0	0.008	0.108	0.007	0.109	0.007	0.108	0.008	0.108
12	0.6	−1	−0.25	−0.004	0.117	−0.005	0.117	−0.004	0.117	−0.005	0.117
13	0.6	0	0.25	0.003	0.106	0.003	0.108	0.002	0.108	0.002	0.107
14	0.6	0	0	−0.014	0.093	−0.014	0.094	−0.014	0.093	−0.013	0.093
15	0.6	0	−0.25	−0.003	0.095	−0.003	0.094	−0.004	0.095	−0.004	0.096
16	0.6	1	0.25	0.003	0.102	0.005	0.102	0.004	0.102	0.004	0.103
17	0.6	1	0	0.005	0.097	0.004	0.098	0.004	0.098	0.004	0.098
18	0.6	1	−0.25	−0.008	0.097	−0.008	0.097	−0.008	0.096	−0.008	0.096
19	0.75	−1	0.25	0.004	0.126	−0.003	0.127	−0.003	0.127	−0.003	0.127
20	0.75	−1	0	0.019	0.105	0.019	0.105	0.019	0.105	0.019	0.105
21	0.75	−1	−0.25	−0.014	0.114	−0.005	0.116	−0.006	0.116	−0.007	0.115
22	0.75	0	0.25	0.004	0.100	−0.002	0.100	−0.002	0.100	−0.001	0.100
23	0.75	0	0	−0.012	0.105	−0.012	0.106	−0.012	0.106	−0.012	0.105
24	0.75	0	−0.25	−0.010	0.093	−0.003	0.093	−0.003	0.094	−0.003	0.094
25	0.75	1	0.25	0.005	0.102	−0.003	0.102	−0.002	0.103	−0.002	0.102
26	0.75	1	0	0.007	0.129	0.006	0.131	0.008	0.131	0.007	0.131
27	0.75	1	−0.25	−0.007	0.095	0.001	0.097	0.001	0.097	0.000	0.097

Note. MH DIF statistics were calculated across 100 replications. MH = Mantel–Haenszel; RMSD = root mean square difference.

Table C4 Bias and RMSD of MH DIF Statistics for Short Test With Large Ability Difference for Focal Group Sample Size of 300

Test version				Simple sum score		Weighted sum score					
						Naive binning		Linear binning		Equipercentile binning	
Item	*a*	*b*	*d*	Bias	RMSD	Bias	RMSD	Bias	RMSD	Bias	RMSD
1	0.48	−1	0.25	0.113	0.465	0.037	0.455	0.044	0.459	0.051	0.458
2	0.48	−1	0	0.005	0.468	−0.076	0.469	−0.076	0.478	−0.079	0.478
3	0.48	−1	−0.25	0.140	0.560	0.039	0.533	0.053	0.534	0.050	0.527
4	0.48	0	0.25	0.121	0.472	0.037	0.465	0.057	0.463	0.059	0.468
5	0.48	0	0	0.033	0.442	−0.045	0.435	−0.041	0.444	−0.040	0.449
6	0.48	0	−0.25	0.094	0.420	−0.012	0.408	−0.013	0.411	−0.015	0.408
7	0.48	1	0.25	0.083	0.470	0.013	0.459	0.009	0.450	0.011	0.450
8	0.48	1	0	0.131	0.512	0.062	0.501	0.068	0.499	0.065	0.502
9	0.48	1	−0.25	0.141	0.492	0.043	0.480	0.045	0.476	0.045	0.472
10	0.6	−1	0.25	−0.054	0.461	−0.036	0.462	−0.025	0.468	−0.033	0.471
11	0.6	−1	0	0.031	0.475	0.029	0.466	0.033	0.476	0.036	0.475
12	0.6	−1	−0.25	−0.060	0.454	−0.086	0.453	−0.080	0.446	−0.083	0.448
13	0.6	0	0.25	−0.032	0.477	−0.020	0.461	−0.027	0.467	−0.029	0.471
14	0.6	0	0	0.077	0.443	0.082	0.438	0.078	0.445	0.081	0.446
15	0.6	0	−0.25	−0.024	0.466	−0.050	0.464	−0.059	0.460	−0.054	0.465
16	0.6	1	0.25	−0.073	0.546	−0.062	0.556	−0.060	0.554	−0.064	0.554
17	0.6	1	0	−0.058	0.507	−0.068	0.512	−0.066	0.497	−0.069	0.492
18	0.6	1	−0.25	0.002	0.531	−0.016	0.539	−0.017	0.535	−0.021	0.536
19	0.75	−1	0.25	−0.131	0.507	−0.024	0.508	−0.028	0.501	−0.039	0.503
20	0.75	−1	0	−0.080	0.541	−0.002	0.537	−0.003	0.536	−0.013	0.540
21	0.75	−1	−0.25	−0.118	0.586	−0.054	0.570	−0.056	0.574	−0.055	0.575
22	0.75	0	0.25	−0.139	0.494	−0.046	0.494	−0.060	0.493	−0.058	0.493
23	0.75	0	0	−0.100	0.476	−0.019	0.478	−0.030	0.479	−0.020	0.477
24	0.75	0	−0.25	−0.115	0.485	−0.057	0.480	−0.063	0.478	−0.068	0.477
25	0.75	1	0.25	−0.112	0.591	0.001	0.581	−0.021	0.579	−0.018	0.579
26	0.75	1	0	−0.190	0.535	−0.105	0.519	−0.115	0.514	−0.113	0.516
27	0.75	1	−0.25	−0.125	0.529	−0.047	0.531	−0.060	0.514	−0.061	0.517

Note. MH DIF statistics were calculated across 100 replications. MH = Mantel–Haenszel; RMSD = root mean square difference.

Table C5 Bias and RMSD MH DIF Statistics for Short Test With Large Ability Difference for Focal Group Sample Size of 1,200

Test version				Simple sum score		Weighted sum score					
						Naive binning		Linear binning		Equipercentile binning	
Item	*a*	*b*	*d*	Bias	RMSD	Bias	RMSD	Bias	RMSD	Bias	RMSD
1	0.48	−1	0.25	0.081	0.243	0.019	0.230	0.028	0.231	0.028	0.231
2	0.48	−1	0	0.080	0.241	0.003	0.227	0.010	0.225	0.012	0.224
3	0.48	−1	−0.25	0.049	0.226	−0.057	0.232	−0.044	0.228	−0.044	0.229
4	0.48	0	0.25	0.044	0.215	−0.021	0.212	−0.014	0.211	−0.014	0.212
5	0.48	0	0	0.083	0.235	−0.004	0.224	0.002	0.224	0.001	0.224
6	0.48	0	−0.25	0.083	0.231	−0.017	0.219	−0.010	0.220	−0.010	0.220
7	0.48	1	0.25	0.011	0.227	−0.057	0.232	−0.048	0.231	−0.047	0.230
8	0.48	1	0	0.070	0.241	−0.010	0.228	−0.003	0.227	−0.003	0.228
9	0.48	1	−0.25	0.064	0.205	−0.040	0.198	−0.029	0.195	−0.030	0.194
10	0.6	−1	0.25	0.021	0.243	0.037	0.245	0.036	0.247	0.036	0.249
11	0.6	−1	0	0.026	0.241	0.027	0.239	0.025	0.241	0.026	0.241
12	0.6	−1	−0.25	−0.005	0.233	−0.028	0.229	−0.029	0.230	−0.028	0.230
13	0.6	0	0.25	0.019	0.237	0.034	0.237	0.033	0.237	0.035	0.237
14	0.6	0	0	−0.023	0.218	−0.025	0.217	−0.025	0.217	−0.027	0.215
15	0.6	0	−0.25	0.021	0.236	−0.001	0.234	−0.001	0.236	0.000	0.236
16	0.6	1	0.25	−0.015	0.237	0.000	0.238	−0.002	0.240	0.000	0.239
17	0.6	1	0	0.002	0.248	−0.007	0.247	−0.002	0.246	−0.004	0.246
18	0.6	1	−0.25	−0.007	0.232	−0.038	0.236	−0.035	0.238	−0.035	0.240
19	0.75	−1	0.25	−0.084	0.230	0.026	0.227	0.020	0.227	0.022	0.227
20	0.75	−1	0	−0.098	0.283	−0.003	0.265	−0.011	0.267	−0.012	0.268
21	0.75	−1	−0.25	−0.095	0.293	−0.007	0.288	−0.016	0.286	−0.015	0.285
22	0.75	0	0.25	−0.089	0.242	0.015	0.228	0.009	0.229	0.011	0.228
23	0.75	0	0	−0.078	0.207	0.014	0.198	0.008	0.198	0.009	0.197
24	0.75	0	−0.25	−0.082	0.271	−0.011	0.261	−0.016	0.264	−0.016	0.264
25	0.75	1	0.25	−0.137	0.291	−0.030	0.256	−0.037	0.256	−0.039	0.255
26	0.75	1	0	−0.079	0.272	0.017	0.265	0.006	0.265	0.006	0.265
27	0.75	1	−0.25	−0.055	0.242	0.013	0.241	0.006	0.240	0.007	0.240

Note. MH DIF statistics were calculated across 100 replications. MH = Mantel–Haenszel; RMSD = root mean square difference.

Table C6 Bias and RMSD MH DIF Statistics for Short Test With Large Ability Difference for Focal Group Sample Size of 4,800

Test version				Simple sum score		Weighted sum score					
						Naive binning		Linear binning		Equipercentile binning	
Item	*a*	*b*	*d*	Bias	RMSD	Bias	RMSD	Bias	RMSD	Bias	RMSD
1	0.48	−1	0.25	0.056	0.112	−0.010	0.097	−0.001	0.097	−0.001	0.097
2	0.48	−1	0	0.065	0.129	−0.018	0.116	−0.007	0.115	−0.007	0.115
3	0.48	−1	−0.25	0.065	0.126	−0.036	0.114	−0.024	0.111	−0.025	0.112
4	0.48	0	0.25	0.079	0.135	0.015	0.110	0.021	0.111	0.020	0.111
5	0.48	0	0	0.070	0.128	−0.013	0.107	−0.005	0.106	−0.005	0.106
6	0.48	0	−0.25	0.078	0.132	−0.023	0.108	−0.013	0.107	−0.012	0.107
7	0.48	1	0.25	0.068	0.137	0.001	0.119	0.009	0.119	0.009	0.119
8	0.48	1	0	0.083	0.137	0.000	0.107	0.010	0.108	0.009	0.108
9	0.48	1	−0.25	0.084	0.135	−0.018	0.105	−0.007	0.105	−0.008	0.104
10	0.6	−1	0.25	−0.021	0.126	−0.004	0.124	−0.006	0.124	−0.006	0.124
11	0.6	−1	0	−0.007	0.124	−0.005	0.123	−0.006	0.123	−0.006	0.123
12	0.6	−1	−0.25	0.001	0.119	−0.023	0.121	−0.019	0.121	−0.020	0.121
13	0.6	0	0.25	−0.018	0.097	−0.005	0.098	−0.006	0.098	−0.006	0.098
14	0.6	0	0	−0.001	0.107	−0.005	0.109	−0.005	0.109	−0.004	0.109
15	0.6	0	−0.25	−0.011	0.106	−0.036	0.113	−0.033	0.113	−0.033	0.113
16	0.6	1	0.25	−0.016	0.117	−0.002	0.119	−0.004	0.118	−0.004	0.118
17	0.6	1	0	−0.015	0.131	−0.020	0.132	−0.018	0.132	−0.018	0.131
18	0.6	1	−0.25	−0.002	0.120	−0.029	0.123	−0.026	0.122	−0.025	0.122
19	0.75	−1	0.25	−0.084	0.152	0.026	0.128	0.017	0.128	0.017	0.128
20	0.75	−1	0	−0.116	0.166	−0.015	0.119	−0.025	0.120	−0.026	0.121
21	0.75	−1	−0.25	−0.128	0.174	−0.046	0.127	−0.055	0.129	−0.055	0.129
22	0.75	0	0.25	−0.075	0.134	0.027	0.113	0.020	0.113	0.021	0.113
23	0.75	0	0	−0.105	0.160	−0.014	0.122	−0.021	0.122	−0.021	0.122
24	0.75	0	−0.25	−0.092	0.155	−0.020	0.127	−0.025	0.129	−0.024	0.128
25	0.75	1	0.25	−0.097	0.152	0.012	0.119	0.001	0.117	0.002	0.118
26	0.75	1	0	−0.107	0.166	−0.009	0.128	−0.019	0.128	−0.018	0.127
27	0.75	1	−0.25	−0.105	0.166	−0.031	0.135	−0.038	0.135	−0.039	0.135

Note. MH DIF statistics were calculated across 100 replications. MH = Mantel–Haenszel; RMSD = root mean square difference.

Suggested citation:

Lu, R., Guo, H., & Dorans, N.J.. (2021). *Robustness of weighted differential item functioning (DIF) analysis: The case of Mantel-Haenszel DIF statistics* (Research Report No. RR-21-12). ETS. https://doi.org/10.1002/ets2.12325

Action Editor: Gautam Puhan

Reviewers: Rebecca Zwick and Skip Livingston