Observed Scores as Matching Variables in Differential Item Functioning Under the One- and Two-Parameter Logistic Models: Population Results

ETS RR–19-06

Hongwen Guo
Neil J. Dorans

December 2019

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

Observed Scores as Matching Variables in Differential Item Functioning Under the One- and Two-Parameter Logistic Models: Population Results

Hongwen Guo & Neil J. Dorans

Educational Testing Service, Princeton, NJ

We derive formulas for the differential item functioning (DIF) measures that two routinely used DIF statistics are designed to estimate. The DIF measures that match on observed scores are compared to DIF measures based on an unobserved ability (theta or true score) for items that are described by either the one-parameter logistic (1PL) or two-parameter logistic (2PL) item response theory (IRT) model. We use two different weighting schemes (uniform weights and item discrimination weights) to construct the observed score matching variable. Our results show that (a) under the 1PL item response model, the observed score-based DIF measures always approximate the true score-based DIF measures very closely; (b) under the 2PL model, when the observed score is the simple sum score, the observed score-based DIF measures underestimate or overestimate the true score-based DIF measures under the null hypothesis of no DIF when the groups are different in ability, and this bias is related to the degree to which the average discrimination parameter underestimates or overestimates the studied discrimination parameter; and (c) under the 2PL model, when the item discrimination weights are used to define the observed score, the observed score-based DIF measures always approximate the true score-based DIF measures very closely. These results will hold for any sets of item responses that are described by either the 1PL or 2PL IRT model.

Keywords Sufficient statistics; IRT; computing algorithm

doi:10.1002/ets2.12243

Differential item functioning (DIF) analysis has been a standard fairness practice conducted in the testing industry since the late 1980s (Dorans, 2013; Holland & Wainer, 1993; Zieky, 1993, 2011; Zwick, 2012). A DIF analysis compares whether a test item functions differently for test takers in a focal group and for test takers of the same ability in a reference group. Specifically, the analysis tests whether, among test takers of the same ability, the probability of correctly responding to an item is different for members of the focal group and for members of the reference group. For dichotomously scored items, the Mantel–Haenszel (MH; Mantel & Haenszel, 1959) common log-odds ratio estimator is widely used in the educational measurement field and other fields to compare a treatment effect on a focal group and a reference group (Agresti, 2002). The MH procedure (i.e., the MH D-DIF statistic), as modified by Holland and Thayer (1988), has became the most widely used methodology and is recognized as the educational testing industry standard (Roussos, Schnipke, & Pashley, 1999; Zwick, 2012).

The MH D-DIF statistic (Holland & Thayer, 1988) is designed to estimate the odds ratio, that is, the ratio of the odds of correct responses for a member of the focal group to the odds of correct responses for a comparable member of a reference group, on a dichotomous item. To define comparable subgroups, a criterion or matching variable is used; typically, it is a measure of the ability that the item and test are designed to assess. The matching variable is used to sort test takers into strata containing individuals of similar ability. In Stratum k, let R_{kf} and R_{kr} be the numbers of test takers in the focal and reference groups, respectively, who answered the studied item correctly (i.e., item response variable $Y = 1$); let W_{kf} and W_{kr} be the numbers of test takers in the focal and reference groups who answered incorrectly, respectively. The odds ratio observed in Stratum k can be computed as

$$O_k = \frac{R_{kr} W_{kf}}{R_{kf} W_{kr}}, \tag{1}$$

Corresponding author: H. Guo, E-mail: hguo@ets.org

and the MH D-DIF statistic is defined as

$$-2.35 \ln \alpha_{\text{MH}} = -2.35 \ln \frac{\sum_k w_k O_k}{\sum_k w_k}, \tag{2}$$

where $w_k = R_{kf} W_{kr} / N_k$ and N_k is the total number of test takers in Stratum k (Dorans & Kulick, 1986; Holland & Thayer, 1988). The most common choice of the matching variable, in practice and in research, is the observed total test score (i.e., the simple sum score). The hope is that when there is no DIF, the MH D-DIF statistic would be close to zero.

As noted in Holland and Thayer (1988; as well as in Donoghue, Holland, & Thayer, 1993; Zwick, 1990; Zwick, Thayer, & Mazzeo, 1997), a simple relation between item parameters and the MH D-DIF statistic exists only when the Rasch model or one-parameter logistic (1PL) model fits the data. Most simulation studies that used the two-parameter logistic (2PL) or the three-parameter logistic (3PL) item response theory (IRT) model to generate data found that MH D-DIF statistics perform differently from values obtained when the true DIF size is based on the IRT latent abilities or the true scores (Donoghue et al., 1993; Roussos et al., 1999; Zwick, 2012).

In the process of evaluating the performance of MH D-DIF statistics, many researchers proposed formulas for computing the true DIF size (the criterion) that the MH D-DIF statistic is designed to measure. Among others, Zwick et al. (1997) and Zwick, Thayer, and Lewis (2000) proposed a general formula based on the IRT latent variable. The true DIF size, $\Delta - \text{DIF}_\theta$, of an item for the 3PL model is defined as (Zwick et al., 2000, p. 234)

$$\Delta - \text{DIF}_\theta = -2.35 \int_\theta \ln \left\{ \frac{P_r(\theta) / Q_r(\theta)}{P_f(\theta) / Q_f(\theta)} \right\} \psi_r(\theta)\, d\theta, \tag{3}$$

where θ is the latent ability in the IRT model, $P_f(\theta) = P_f(Y = 1 | \theta)$ and $P_r(\theta) = P_r(Y = 1 | \theta)$ are the item response functions for the focal and reference groups for the studied item Y, respectively, $Q(\theta) = 1 - P(\theta)$, and $\psi_r(\theta)$ is the density function of the latent ability of the reference group. That is, $\Delta - \text{DIF}_\theta$ is the expected value of the log odds ratio based on θ in the reference group. Roussos et al. (1999) discussed an asymptotic-based true DIF size of the MH D-DIF statistic, which results in the same formula proposed by Spray and Miller (1992). Their suggested formula of $\Delta - \text{DIF}_\theta$ was derived from the MH statistic by letting the sample size and number of items tend to infinity. However, this general formula in Roussos et al. (1999) has an additional integrand that is related to the sample sizes of the reference and focal groups. Roussos et al. (1999) also observed that, for 3PL models, the bias between MH D-DIF statistics and the true DIF size increased as item difficulty increased, especially for items with larger item discrimination parameters. Zwick (1990) observed that, under the 2PL and 3PL models, identity of item response functions for the two populations did not imply that the MH null hypothesis would be satisfied. Simulation studies have demonstrated the efficacy of MH D-DIF statistics matching on total score for 3PL items in a wide variety of situations (e.g., Allen & Donoghue, 1996; Chang, Mazzeo, & Roussos, 1995; Roussos & Stout, 1996; Shealy & Stout, 1993).

The MH D-DIF statistic (Holland & Thayer, 1988) was adapted from the MH procedure that measures the expected value of log odds ratio based on observed score X, that is,

$$\Delta - \text{DIF}_X = -2.35 \sum_x \ln \left\{ \frac{P_r(x) / Q_r(x)}{P_f(x) / Q_f(x)} \right\} g_r(x), \tag{4}$$

where x is the observed simple sum score, $P(x) = (Y = 1 | X = x)$, $Q(x) = 1 - P(x)$, and $g_r(x)$ is the probability distribution of x for the reference group.[1] It is important to evaluate the conditions under which $\Delta - \text{DIF}_X$ and $\Delta - \text{DIF}_\theta$ are expected to be different from each other.[2]

Another widely used DIF statistic for dichotomous items is the Dorans–Kulick standardization method based on P^+, the observed item difficulty or mean item score (Dorans & Kulick, 1986). It is a weighted difference of conditional P^+ values between the focal group and the reference group for an item:

$$\text{STD P-DIF} = \sum_k w_{kf} P^+_{kf} - \sum_k w_{kf} P^+_{kr}, \tag{5}$$

where, at the kth level of the stratum/matching variable (typically the total score), w_{fk} is the proportion of focal group members in strata k and $P^+_{kf} = R_{kf} / \left(R_{kf} + W_{kf} \right)$ and $P^+_{kr} = R_{kr} / \left(R_{kr} + W_{kr} \right)$ are the mean item scores for the focal and reference groups, respectively. Similarly, the criterion $P - \text{DIF}_\theta$ is calculated by matching the item scores on the latent variable θ (Chang et al., 1995; Shealy & Stout, 1993),

$$P - \mathrm{DIF}_{\theta} = \int_{\theta} \left[P_f(\theta) - P_r(\theta)\right] \psi_f(\theta)\, d\theta, \tag{6}$$

which is the expected difference in the response function $P(\theta)$ with respect to the focal group, whereas the STD P-DIF statistic measures the expected difference in the response function $P(X)$ with respect to the focal group, that is,

$$P - \mathrm{DIF}_{X} = \sum_{x} \left[P_f(x) - P_r(x)\right] g_f(x), \tag{7}$$

where $\psi_f(\theta)$ and $g_f(x)$ are the density function and the probability distribution of θ and x, respectively, for the focal group.

Both significance tests and effect sizes exist for the two observed score-based DIF statistics: MH D-DIF and STD P-DIF. Dorans and Holland (1993) discussed the two DIF statistics and their commonalities and differences.

Numerous investigations of the differences between observed score-based DIF statistics and latent ability or true score-based criteria were conducted in simulations, and the differences were often interpreted as results affected by randomness and limited sample sizes, particularly for cases where the generating probabilities are small. In this report, to avoid these issues, we analytically investigated when the observed score-based DIF measures, $\Delta - \mathrm{DIF}_X$ and $P - \mathrm{DIF}_X$, are different from the true score-based DIF measures, $\Delta - \mathrm{DIF}_\theta$ and $P - \mathrm{DIF}_\theta$. We also evaluated what factors impact the differences between the pairs $\Delta - \mathrm{DIF}_X$ versus $\Delta - \mathrm{DIF}_\theta$ and $P - \mathrm{DIF}_X$ versus $P - \mathrm{DIF}_\theta$. We examined two weighting schemes for constructing the matching variable X: uniform weights (i.e., X is the simple sum score) and item discrimination weights (i.e., X is the weighted sum of item scores with weights equal to item discrimination parameters). Our analytical results show that matching on sufficient statistics of the latent ability could significantly improve the performance of these DIF statistics.

Study Designs

Matching Variables

Because most testing programs report a simple sum score or a transformation of it, it makes sense to study the performance of the simple sum score as a matching variable. In addition, the use of the simple sum scores as the matching variable has been justified by Holland and Thayer (1988) for the 1PL model. However, because of widely observed bias under the null hypothesis of no DIF when the 2PL or the 3PL models generate the data and when the focal and reference groups differ in ability, we also investigated DIF measures that matched on a weighted sum score, where the weights equal the item discrimination parameters.

Under the 1PL model, the simple sum score is a sufficient statistic for the latent ability θ (Holland & Thayer, 1988; Lord, 1980). Because the item discrimination parameters are the same under 1PL, weighting by the item discrimination parameter is the same as using equal weights. Matching on the simple sum score is the best approximation to matching on θ (Holland & Thayer, 1988), and hence we expect that the observed score-based DIF measures perform similarly to the true score-based DIF measures no matter whether the group difference exists or not.

Under the 2PL model, both the weighted sum (with weights equal to item discrimination parameters; Lord, 1980, p. 57) and the item response vector are sufficient statistics of the latent ability θ, and the posterior mean of θ is a strictly increasing function of the weighted sum (refer to Appendix B). However, the simple sum score loses information concerning θ. When matching variables are constructed by the two different weighting schemes, we expect differences in the observed score-based DIF measures associated with these different matching values; results based on the weighted sum score are expected to be closer to the true score-based DIF measures than those obtained with the simple sum score.

Test Design

As in a typical DIF study, in each of the following DIF analyses, only the studied item has a difficulty parameter for the focal group different from that for the reference group; the rest of the items stay the same for the two groups.

We reviewed previous study designs (Camilli & Shepard, 1994; Chang et al., 1995; Penfield, 2007; Zwick et al., 1997) and decided to manipulate the following factors:

- The item difficulty b was set to be −1, 0, or 1 to represent an easy, medium difficult, or hard item, respectively.

- The item discrimination parameter a was set to be 0.60 in the Rasch model and to be 0.48, 0.60, or 0.75 in the 2PL model, to approximate the lower quartile, median, and upper quartile of a log-normal distribution derived from a large-scale standardized test, respectively.
- The difference in item difficulties for the studied item between the focal and reference groups $d = b_f - b_r$ was set to be −0.25, 0, or 0.25 (the item discrimination parameter did not differ).
- The reference group ability followed a standard normal distribution $N(0.5, 1)$, and the focal group ability followed either $N(0.5, 1)$ or $N(-0.5, 1)$ to represent two cases: when there was no group difference and when there was a big group difference.
- Use of three levels of item difficulty and three levels of discrimination (for the 2PL model) resulted in nine distinct classes of items. For each DIF analysis, the studied item can have three levels of item difficulty difference between the focal and reference groups, so there are $27 = 3 \times 3 \times 3$ possible DIF cases on these nine classes. The test length L was set to be 27, 54, or 108 to represent a short test, a medium-length test, or a long test, respectively.

Note that, to simplify our illustrations, we only have three levels of the item discrimination parameters instead of L distinct item discrimination parameters.

Computation Related to Observed Score Distributions

The following formulas are generic and applicable to both parametric IRT models (such as the 1PL, 2PL, and 3PL models) and nonparametric IRT models for dichotomous responses.

Uniformly Weighted Simple Sum Scores

To obtain the observed score distribution for the uniformly weighted sum scores X, we used the iterative method in Lord and Wingersky (1984) and Kolen and Brennan (2014). For a test taker of θ_i, let p_{ij} be the probability of obtaining a correct answer on Item j and $f_j(x|\theta_i)$ the distribution of number-correct scores over the first j items. Define $f_1(x=1|\theta_i) = p_{i1}$, $f_1(x=0|\theta_i) = 1 - p_{i1}$. For $j > 1$, the recursion formula is as follows (Kolen & Brennan, 2014):

$$f_j\left(x|\theta_i\right) = \begin{cases} f_{j-1}\left(x|\theta_i\right)\left(1-p_{ij}\right), & x = 0, \\ f_{j-1}\left(x|\theta_i\right)\left(1-p_{ij}\right) + f_{j-1}\left(x-1|\theta_i\right)p_{ij}, & 0 < x < j, \\ f_{j-1}\left(x-1|\theta_i\right)p_{ij}, & x = j. \end{cases} \tag{8}$$

Marginal Distribution of Observed Scores

Assume we have the conditional distribution of the simple sum on the test for a given ability $f(x|\theta)$ from Equation 8; then, the observed score distribution is

$$g(x) = \int_\theta f(x|\theta)\,\psi(\theta)\,d\theta, \tag{9}$$

where $\psi(\theta)$ is the ability density distribution.

Conditional Probability Given Observed Scores

To compute Equation 4, we need to compute $p_j(x) = P(Y_j|X=x)$, where Y_j is the dichotomous response of Item j. Note that

$$\begin{aligned} P\left(Y_j = 0|X = x\right) &= \int_\theta P\left(Y_j = 0|X = x, \theta\right)\psi(\theta)\,d\theta \\ &= \int_\theta \frac{P\left(Y_j = 0, X = x|\theta\right)}{f(x|\theta)}\psi(\theta)\,d\theta \\ &= \int_\theta \frac{P\left(X = x|Y_j = 0, \theta\right) \times P\left(Y_j = 0|\theta\right)}{f(x|\theta)}\psi(\theta)\,d\theta, \end{aligned} \tag{10}$$

where $f(x|\theta)$ can be obtained from Equation 8 and $P(Y_j = 0|\theta)$ can be calculated from the IRT model. To obtain the term $P(X = x|Y_j = 0, \theta)$, we use the recursion method in Equation 8 to obtain $f_{-j}(x|\theta)$, the score distribution of $X = x$ given θ without Item j on the test, then let $P(X = x|Y_j = 0, \theta) = f_{-j}(x|\theta)$.

Density of True Scores

The true score $T_X(\theta)$ corresponding to θ is obtained as the sum of expected item scores on the test form X; that is,

$$T_X(\theta) = \sum_{j=1}^{J} P\left(Y_j = 1|\theta\right). \tag{11}$$

For simplicity, we use $T(\theta)$ to denote $T_X(\theta)$. The density distribution of $T(\theta)$ equals

$$\psi(\theta) \times \frac{d\theta}{dT(\theta)}.$$

In the case that the 2PL model is used, the true score density is

$$\psi(\theta) \times \left\{ \sum_{j=1}^{J} \frac{D_0 a_j \exp\left[D_0 a_j\left(\theta - b_j\right)\right]}{\left\{1 + \exp\left[D_0 a_j\left(\theta - b_j\right)\right]\right\}^2} \right\}^{-1},$$

where $D_0 = 1.702$ and $\theta = T^{-1}(\theta)$, the inverse function in Equation 11.

For a given a true score $T(\theta)$, the Newton–Raphson method is used to find the corresponding θ (Kolen & Brennan, 2014).

Calculation of Test Reliability

The test reliability ρ is defined as

$$\rho^2 = \frac{\sigma_T^2}{\sigma_X^2} = 1 - \frac{\sigma_E^2}{\sigma_X^2} = 1 - \frac{\int \sigma^2(X|\theta)\,\psi(\theta)\,d\theta}{\sigma_X^2}, \tag{12}$$

where σ_T, σ_X, and σ_E are the standard deviation of the true score, observed score, and measurement error, respectively, and $\sigma(X|\theta)$ is the conditional standard deviation of X given θ. Both $\sigma^2(X|\theta)$ and σ_X^2 can be obtained from Equations 8 and 9 by definition (Kolen, Zeng, & Hanson, 1996), respectively.

Sum Scores Weighted by Item Discrimination Parameters

To compute the observed score-based DIF measures with weighted sum scores, we used the convolution of probability distribution technique: If Z is the sum of two independent variables X and Y, then

$$P(Z = z) = \sum_k P(X = k) \times P(Y = z - k).$$

Similar algorithms were also shown in Thissen, Pommerich, Billeaud, and Willams (1995) and Haberman (2013).

To illustrate, we now study the last item on the short test (i.e., $L = 27$), in which the first nine, the second nine, and the third nine items have the item discrimination parameter $a = 0.48$, 0.60, and 0.75, respectively. Using Equation 8, we computed the conditional sum score distributions $f_1(x|\theta)$, $f_2(x|\theta)$, and $f_3(x|\theta)$ given θ for the first nine, second nine, and last nine items that have the same a parameter, respectively.

To obtain the conditional weighted sum score distribution $f_{1+2}(x_w|\theta)$ given θ for the first 18 items, we used the following iterative formulas. Let z_{1+2} be the possible weighted sum scores on the 18 items; let s_x, s_y, and s_z be the possible sum scores of the first nine items, the second nine items, and the third nine items, respectively; and let $w_1 = .48$, $w_2 = .60$, and $w_3 = .75$. Then,

$$z_{1+2} = w_1 s_x + w_2 s_y,$$

$$f_{1+2}\left(x_w = z_{1+2}|\theta\right) = \sum_k f_1\left(s_x = k|\theta\right) \times f_2\left(s_y = \left(z_{1+2} - w_1 k\right)/w_2|\theta\right).$$

Similarly, let z_{1+2+3} be the possible weighted sum scores of the 27 items, and let $f_{1+2+3}(x_w|\,\theta)$ be the conditional weighted sum score distribution given θ:

$$z_{1+2+3} = z_{1+2} + w_3 s_z$$

$$f_{1+2+3}\left(x_w = z_{1+2+3}|\theta\right) = \sum_k f_{1+2}\left(z_{1+2} = k|\theta\right) \times f_3\left(w_3 s_z = z_{1+2} - k|\theta\right).$$

Other related distributions of weighted sum scores can be computed based on methods in the section Uniformly Weighted Simple Sum Scores in a similar fashion. For our simple test design with only three unique values of item discrimination parameters (i.e., $a = 0.48$, 0.60, and 0.75), the numbers of possible weighted scores are 414, 959, and 2,052, respectively, for the test of length $L = 27$, 54, and 108. Even in the case of only three unique discrimination parameters, these observed score-based DIF measures are computationally intensive.

Analytical Results

Note that under both 1PL and 2PL models, for the uniform DIF case (i.e., the item discrimination parameter is the same, but the item difficulty parameters are different for the focal and reference groups), we have

$$-2.35 \times \ln\left\{P_r(\theta)\,Q_f(\theta)/\left[Q_r(\theta)\,P_f(\theta)\right]\right\} = -4a\left(b_f - b_r\right) \tag{13}$$

for the studied item, which is independent of θ, and we expect $\Delta - \mathrm{DIF}_\theta = -4a(b_f - b_r)$ as well.

Based on results in Cressie and Holland (1981), Holland and Thayer (1988) demonstrated that, for the uniform DIF case under the 1PL model,

$$-2.35 \times \ln\left\{P_r(X)\,Q_f(X)/\left[Q_r(X)\,P_f(X)\right]\right\} = -4a\left(b_f - b_r\right), \tag{14}$$

where the matching variable X is the simple sum score of all items including the studied item, when none of the other items have DIF.

Similarly, based on Cressie and Holland (1981), we can readily derive that, under the 2PL model, if we replace the simple sum score by the weighted sum score with weights equal to the item discrimination parameters, Equation 14 still holds. Therefore, theoretically, $\Delta - \mathrm{DIF}_X$ is independent of the test length L and the group ability difference when the matching variable is a sufficient statistic for the latent ability (refer to Appendix B). However, $P - \mathrm{DIF}_\theta$ is a function of the group difference by Definition 6, and $P - \mathrm{DIF}_X$ depends on both group difference and test length, as shown in Equation 7.

To evaluate the differences between the observed score- and true score-based DIF measures, we present the numerical results in the following tables. To obtain the numerical values of these DIF measures, the integration in Equation 9 has to be approximated by

$$g(x) \sim \sum_i f\left(x|\theta_i\right) w_i, \tag{15}$$

where θ_i are the nodes and w_i are the corresponding weights. There are several ways to approximate the integration via Gaussian quadratures and adaptive quadratures (Haberman, 2013). We used the simple and widely used naive approach in the following illustration: 41 equally spaced θ_i nodes on the interval of $[-4, 4]$ and their corresponding weights.[3]

The tables to follow contain numerical approximations of the DIF measures. For each row in any one of the tables, only the studied item (which is the item on that row) is different between the focal group and the reference group; the rest of the items in that table are the same (i.e., $d = b_f - b_r = 0$) for the two groups.

One-Parameter Logistic Model

Under the 1PL model, we expect that $\Delta - \mathrm{DIF}_\theta$ and $\Delta - \mathrm{DIF}_X$ will be closer and closer to each other as the test length increases, regardless of the manipulated factors in the test design from Equation 14. However, there are no a priori expectations about the differences between $P - \mathrm{DIF}_\theta$ and $P - \mathrm{DIF}_X$. Table 1 displays the DIF measures based on the simple sum score and the true ability, where $D = \mu_r - \mu_f$ is the mean ability difference between the reference and focal groups; $L = 27$

Table 1 Differential Item Functioning Measures Under the One-Parameter Logistic Model for a Short Test ($L = 27$)

Item	a	b_r	d	$\Delta-\mathrm{DIF}_\theta$	$\Delta-\mathrm{DIF}_X$	$d(\Delta)$	$P-\mathrm{DIF}_\theta$	$P-\mathrm{DIF}_X$	$d(P)$
$D = 0$									
1	0.60	−1	0.25	−0.5992	−0.5957	0.0036	−0.0393	−0.0380	0.0014
2	0.60	−1	0.00	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
3	0.60	−1	−0.25	0.5993	0.5957	−0.0036	0.0352	0.0339	−0.0012
4	0.60	0	0.25	−0.5992	−0.5957	0.0036	−0.0512	−0.0491	0.0021
5	0.60	0	0.00	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
6	0.60	0	−0.25	0.5993	0.5957	−0.0036	0.0493	0.0473	−0.0020
7	0.60	1	0.25	−0.5992	−0.5957	0.0036	−0.0493	−0.0470	0.0023
8	0.60	1	0.00	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
9	0.60	1	−0.25	0.5993	0.5957	−0.0036	0.0512	0.0489	−0.0023
$D = 1$									
1	0.60	−1	0.25	−0.5992	−0.5957	0.0036	−0.0512	−0.0489	0.0023
2	0.60	−1	0.00	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
3	0.60	−1	−0.25	0.5993	0.5957	−0.0036	0.0493	0.0470	−0.0023
4	0.60	0	0.25	−0.5992	−0.5957	0.0036	−0.0493	−0.0473	0.0020
5	0.60	0	0.00	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
6	0.60	0	−0.25	0.5993	0.5957	−0.0036	0.0512	0.0491	−0.0021
7	0.60	1	0.25	−0.5992	−0.5957	0.0036	−0.0352	−0.0339	0.0012
8	0.60	1	0.00	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
9	0.60	1	−0.25	0.5993	0.5957	−0.0036	0.0393	0.0379	−0.0014

is the test length, b_r is the studied item difficulty in the reference group, $d = b_f - b_r$ is the studied item difficulty difference between the focal and reference groups, $d(\Delta) = \Delta - \mathrm{DIF}_X - \Delta - \mathrm{DIF}_\theta$, and $d(P) = P - \mathrm{DIF}_X - P - \mathrm{DIF}_\theta$. For this short test, the reliability is .84.

In the 1PL model, the item discrimination was chosen to be $a = .60$, the median discrimination of items for the large-scale assessment to which we referred earlier. In Table 1, the DIF measures of nine cases are reported because we have only nine distinct DIF cases for the fixed a parameter.

In Table 1, the DIF measures between no ability difference ($D = 0$) and ability difference of $D = 1$ for the 1PL case are very similar. For a short test of 27 items, both observed score-based DIF measures are zero under the null hypothesis of no DIF. The differences $d(\Delta) = \Delta - \mathrm{DIF}_X - \Delta - \mathrm{DIF}_\theta$ and $d(P) = P - \mathrm{DIF}_X - P - \mathrm{DIF}_\theta$ are very small (to the third decimal place) when $d = b_f - b_r = .25$. These differences are due to errors caused by numerical approximation of the analytical results and the limited test length; they decrease as the test length increases (see Table A1, which contains results for test lengths of 54 and 108). In addition, compared to $\Delta - \mathrm{DIF}$, which is not affected by item difficulty b_r or ability difference $D = \mu_f - \mu_r$, the $P - \mathrm{DIF}$ measures and their difference $d(P)$ decrease in magnitude when the relative item difficulty $b^* = b_r + d - \mu_f$ increases in absolute value. Also observe that the observed score-based DIF measures are regressed to zero. Consequently, observed score-based DIF measures will tend to be less likely to detect DIF than their true score-based counterparts. For example, for the first three cases in Table 1, $\Delta - \mathrm{DIF}_\theta = -.5992, 0, .5993$ and $\Delta - \mathrm{DIF}_X = -.5957, 0, .5957$. The latter set is slightly closer to 0 in absolute value.

Because of the short test length, with a reliability of .84, there are differences between the observed and true score density distributions (as seen in Figure A1). In contrast, the conditional log odds ratio $\ln\{P_r(x)Q_f(x)/[Q_r(x)P_f(x)]\}$ for a given observed score X is nearly constant across scores (an assumption used in the MH D-DIF procedure; see Figure A2); the differences between the odds ratios based on X and on $-4a(b_f - b_r)$ in Equation 13 are negligible.

The test reliability is .91 for $L = 54$ and .95 for $L = 108$ for the longer tests. As mentioned earlier, as the test becomes longer, under the 1PL model, the difference between the true score-based and observed score-based DIF measures converges to zero (as seen in Table A1). These results reflect the increase in test score reliabilities from .84 ($L = 27$) to .95 ($L = 108$).

Two-Parameter Logistic Model With the Simple Sum Scores

We examine the results for the 2PL model when there is no group ability difference and when there is large group difference, respectively, in this section. When the item discrimination parameters are known, the 2PL IRT model is what Verhelst and Glas (1995) referred to as a one-parameter logistic model (OPLM).

Table 2 Differential Item Functioning Measures When $D = 0$ Under the Two-Parameter Logistic Model for a Short Test ($L = 27$)

Item	a	b_r	d	$\Delta - \text{DIF}_\theta$	$\Delta - \text{DIF}_X$	$d(\Delta)$	$P - \text{DIF}_\theta$	$P - \text{DIF}_X$	$d(P)$
1	0.48	−1	0.25	−0.4794	−0.4730	0.0064	−0.0359	−0.0333	0.0026
2	0.48	−1	0.00	−0.0000	−0.0000	0.0000	0.0000	0.0000	0.0000
3	0.48	−1	−0.25	0.4794	0.4732	−0.0062	0.0331	0.0310	−0.0021
4	0.48	0	0.25	−0.4794	−0.4723	0.0071	−0.0438	−0.0392	0.0045
5	0.48	0	0.00	−0.0000	−0.0000	0.0000	0.0000	0.0000	0.0000
6	0.48	0	−0.25	0.4794	0.4724	−0.0070	0.0425	0.0383	−0.0042
7	0.48	1	0.25	−0.4794	−0.4724	0.0070	−0.0425	−0.0382	0.0043
8	0.48	1	0.00	−0.0000	−0.0000	0.0000	0.0000	0.0000	0.0000
9	0.48	1	−0.25	0.4794	0.4723	−0.0071	0.0438	0.0391	−0.0046
10	0.60	−1	0.25	−0.5992	−0.5959	0.0033	−0.0393	−0.0379	0.0014
11	0.60	−1	0.00	−0.0000	−0.0000	0.0000	0.0000	0.0000	0.0000
12	0.60	−1	−0.25	0.5993	0.5959	−0.0033	0.0352	0.0339	−0.0013
13	0.60	0	0.25	−0.5992	−0.5959	0.0033	−0.0512	−0.0490	0.0022
14	0.60	0	0.00	−0.0000	−0.0000	0.0000	0.0000	0.0000	0.0000
15	0.60	0	−0.25	0.5993	0.5959	−0.0033	0.0493	0.0471	−0.0022
16	0.60	1	0.25	−0.5992	−0.5959	0.0033	−0.0493	−0.0469	0.0023
17	0.60	1	0.00	−0.0000	−0.0000	0.0000	0.0000	0.0000	0.0000
18	0.60	1	−0.25	0.5993	0.5959	−0.0033	0.0512	0.0488	−0.0024
19	0.75	−1	0.25	−0.7491	−0.7389	0.0101	−0.0417	−0.0407	0.0010
20	0.75	−1	0.00	−0.0000	−0.0000	0.0000	0.0000	0.0000	0.0000
21	0.75	−1	−0.25	0.7491	0.7398	−0.0093	0.0361	0.0344	−0.0017
22	0.75	0	0.25	−0.7491	−0.7359	0.0132	−0.0587	−0.0600	−0.0012
23	0.75	0	0.00	−0.0000	−0.0000	0.0000	0.0000	0.0000	0.0000
24	0.75	0	−0.25	0.7491	0.7364	−0.0127	0.0559	0.0566	0.0007
25	0.75	1	0.25	−0.7491	−0.7362	0.0129	−0.0559	−0.0563	−0.0004
26	0.75	1	0.00	−0.0000	−0.0000	0.0000	0.0000	0.0000	0.0000
27	0.75	1	−0.25	0.7491	0.7357	−0.0133	0.0587	0.0598	0.0010

When $D = \mu_r - \mu_f = 0$

Table 2 shows the DIF measures for each studied item when $D = 0$ of the short test ($L = 27$). In this case, the DIF measures are zero under the null hypothesis; under the alternate hypothesis, $\Delta - \text{DIF}_X$ is slightly regressed to zero compared to $\Delta - \text{DIF}_\theta$, and $P - \text{DIF}_X$ is slightly regressed to zero compared to $P - \text{DIF}_\theta$, except for $a = .75$. All the DIF measures increase in magnitude as a increases. The difference $d(\Delta) = \Delta - \text{DIF}_X - \Delta - \text{DIF}_\theta$ increases in absolute value as $(a - .60)$ increases in absolute value, but the difference $d(P) = P - \text{DIF}_X - P - \text{DIF}_\theta$ decreases in absolute value as a increases. In addition, for each fixed a, the $P - \text{DIF}$ measures decrease in absolute value when the relative item difficulty $b^* = b_r + d - \mu_f$ increases in absolute value. Although the difference between the observed score- and the true score-based DIF measures is larger in absolute value compared to the 1PL case as a deviates from .60 (the average discrimination parameter), it may not be large enough to be of practical concern.

Figure 1 shows the conditional log odds ratio $\ln\{P_r(x)Q_f(x)/[Q_r(x)P_f(x)]\}$ (represented by the triangles) for a given simple sum score $X \in \{0, 1, \cdots, 27\}$ for various DIF cases. For example, the top left panel shows the conditional log odds ratio when $b_r = -1$ and $d = b_f - b_r = .25$; the three curves made of triangles in the middle of the panel from top to bottom correspond to conditional log odds ratio for $a = .48$, .60, and .75, respectively. The three solid lines stand for the constants derived from $-4a(b_f - b_r)$ in Equation 13. It shows clearly that the log odds ratio $\ln\{P_r(x)Q_f(x)/[Q_r(x)P_f(x)]\}$ is not constant across scores (a clear violation of the uniform DIF assumption used in the MH D-DIF procedure) when DIF exists. However, the differences between the log odds ratios based on X and the constant $-4a(b_f - b_r)$ are relatively small (particularly for items with $a = .60$).

Tables A2 and A3 show the DIF measures for the longer tests when $D = 0$ under the 2PL model. There we observe that as the test gets longer, neither are the conditional odds ratios constant across scores, similar to the results for $L = 27$. The differences between the observed score-based DIF measures and the true score-based DIF measures exist, but with a smaller magnitude. It appears that the difference might not converge to zero for any reasonably long tests when $a \neq 0.6$.

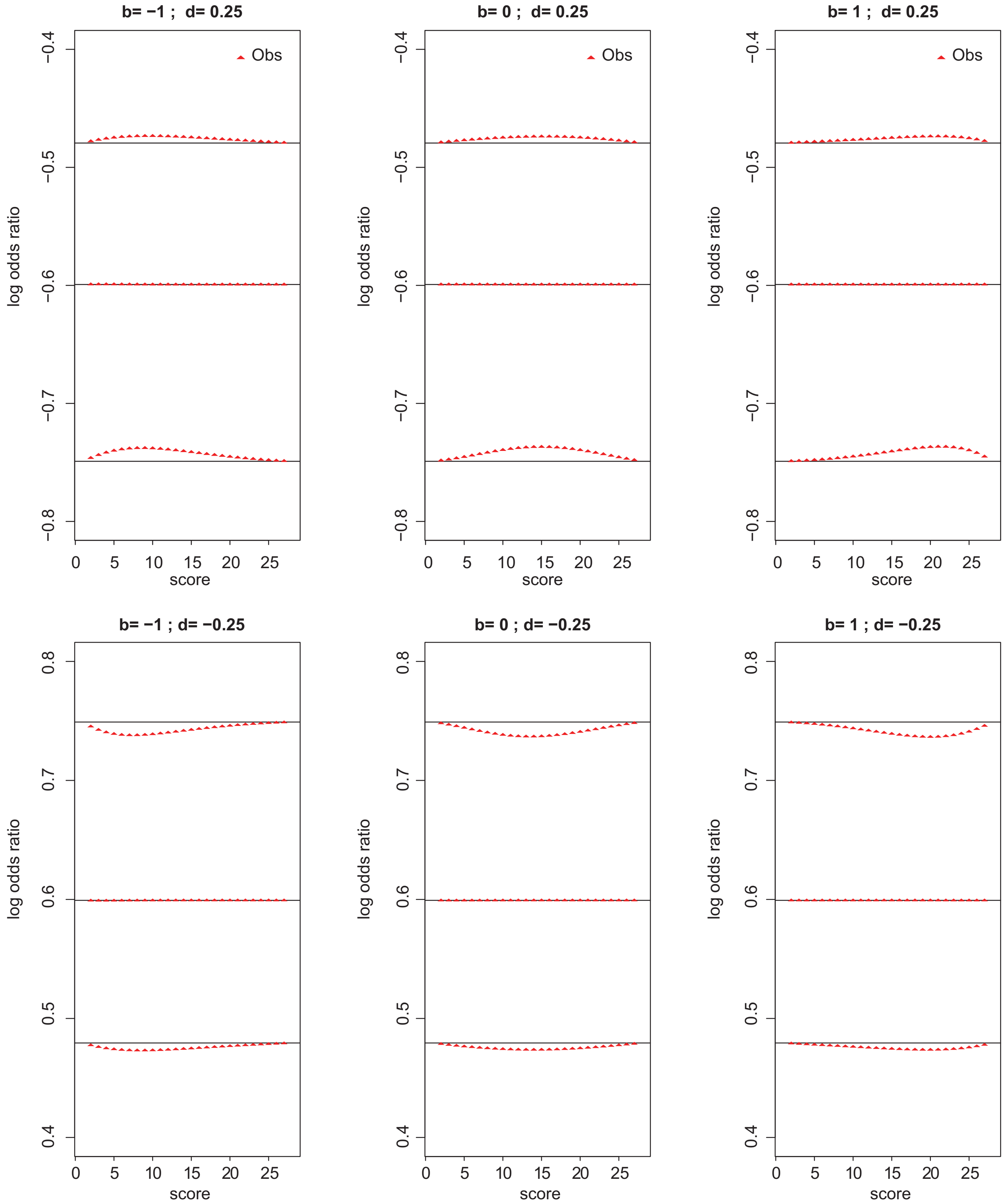

Figure 1 Log odds ratios based on X when $D = 0$ under the two-parameter logistic model for a short test ($L = 27$). Within each panel, the three curves from top to bottom stand for items of $a = .48$, .60, and .75, respectively.

Table 3 Differential Item Functioning Measures When $D = 1$ Under the Two-Parameter Logistic Model for a Short Test ($L = 27$)

Item	a	b_r	d	$\Delta - \text{DIF}_\theta$	$\Delta - \text{DIF}_X$	$d(\Delta)$	$P - \text{DIF}_\theta$	$P - \text{DIF}_X$	$d(P)$
1	0.48	−1	0.25	−0.4794	0.0424	0.5218	−0.0438	0.0040	0.0477
2	0.48	−1	0.00	−0.0000	0.5156	0.5156	0.0000	0.0431	0.0431
3	0.48	−1	−0.25	0.4794	0.9891	0.5097	0.0425	0.0813	0.0388
4	0.48	0	0.25	−0.4794	0.0470	0.5264	−0.0425	0.0034	0.0460
5	0.48	0	0.00	−0.0000	0.5193	0.5193	0.0000	0.0418	0.0418
6	0.48	0	−0.25	0.4794	0.9918	0.5124	0.0438	0.0811	0.0373
7	0.48	1	0.25	−0.4794	0.0495	0.5289	−0.0331	0.0026	0.0357
8	0.48	1	0.00	−0.0000	0.5215	0.5215	0.0000	0.0337	0.0337
9	0.48	1	−0.25	0.4794	0.9935	0.5141	0.0359	0.0671	0.0311
10	0.60	−1	0.25	−0.5992	−0.5654	0.0338	−0.0512	−0.0465	0.0047
11	0.60	−1	0.00	−0.0000	0.0305	0.0305	0.0000	0.0023	0.0023
12	0.60	−1	−0.25	0.5993	0.6265	0.0272	0.0493	0.0492	−0.0000
13	0.60	0	0.25	−0.5992	−0.5652	0.0341	−0.0493	−0.0451	0.0042
14	0.60	0	0.00	−0.0000	0.0307	0.0307	0.0000	0.0020	0.0020
15	0.60	0	−0.25	0.5993	0.6267	0.0274	0.0512	0.0510	−0.0002
16	0.60	1	0.25	−0.5992	−0.5652	0.0340	−0.0352	−0.0324	0.0028
17	0.60	1	0.00	−0.0000	0.0307	0.0307	0.0000	0.0015	0.0015
18	0.60	1	−0.25	0.5993	0.6266	0.0274	0.0393	0.0393	0.0001
19	0.75	−1	0.25	−0.7491	−1.3108	−0.5618	−0.0587	−0.1051	−0.0463
20	0.75	−1	0.00	−0.0000	−0.5726	−0.5726	0.0000	−0.0453	−0.0453
21	0.75	−1	−0.25	0.7491	0.1665	−0.5826	0.0559	0.0111	−0.0448
22	0.75	0	0.25	−0.7491	−1.3149	−0.5658	−0.0559	−0.1040	−0.0481
23	0.75	0	0.00	−0.0000	−0.5789	−0.5789	0.0000	−0.0475	−0.0475
24	0.75	0	−0.25	0.7491	0.1572	−0.5918	0.0587	0.0124	−0.0464
25	0.75	1	0.25	−0.7491	−1.3199	−0.5709	−0.0361	−0.0659	−0.0298
26	0.75	1	0.00	−0.0000	−0.5826	−0.5826	0.0000	−0.0316	−0.0316
27	0.75	1	−0.25	0.7491	0.1540	−0.5951	0.0417	0.0089	−0.0327

When $D = \mu_r - \mu_f = 1$

Table 3 displays the DIF measures when $D = 1$ under the 2PL model for the short test ($L = 27$).

Observe that when the group difference is large (i.e., $D = 1$), the observed score-based DIF measures are not zero under the null hypothesis of no DIF, even for $a = .6$. The observed score-based DIF measures exhibit positive bias when $a = .48$ and negative bias when $a = .75$. All the observed score-based DIF measures are quite different from the true score-based DIF measures; the differences are to the first decimal places and the second decimal places for $\Delta - \text{DIF}$ and $P - \text{DIF}$, respectively.

Figure 2 shows $d(\Delta) = \Delta - \text{DIF}_X - \Delta - \text{DIF}_\theta$ (the upper panel) and $d(P) = P - \text{DIF}_X - P - \text{DIF}_\theta$ (the lower panel), and the differences are relatively large compared to previous ones, particularly for $a \neq .60$.

Figure 3 shows the density functions of the true score $T(\theta)$ for the focal and reference groups (labeled as Δ and $^\circ$, respectively) and the observed score X for the focal and reference groups (labeled as X and $+$, respectively) for $\mu_f = -.25$ and $\mu_r = .25$. Because of the short test length, there are differences between the observed and the true score density distributions.

Figure 4 shows the conditional log odds ratios for given values of $X \in \{0, 1, \cdots, 27\}$ (the triangle curves). For example, the top left panel shows the conditional odds ratio $\ln\{P_r(x)Q_f(x)/[P_f(x)Q_r(x)]\}$ for given $X = x$ (observed score; labeled by triangles) when $b_r = -1$, $d = b_f - b_r = .25$ and $a = .48$ (top line), .60 (middle line), and .75 (bottom line), respectively. The solid lines are the constants $-4a(b_f - b_r)$ in Equation 13 for the three studied items. Observe again that, under the 2PL model, the observed score-based odds ratio is obviously not constant when uniform weights are used for the matching variable. When $a \neq .60$, the observed score-based log odds ratio deviates substantially from the constant $-4a(b_f - b_r)$ in Equation 13.

Tables A4 and A5 show the observed score-based DIF measures and their true score-based versions for longer tests when $D = 1$ under the 2PL model. As with the short test, the observed score-based DIF measures indicate DIF under the null hypothesis of no DIF. The observed score-based DIF measures exhibit positive bias when $a = .48$ and negative bias when $a = .75$. It appears that they will not converge to zero for any practical test with a finite length. In addition, when

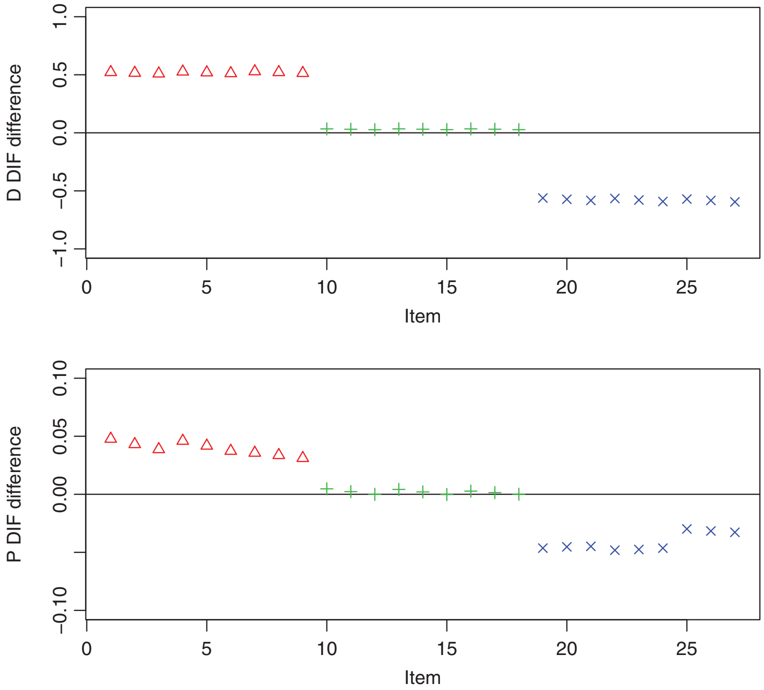

Figure 2 Differences in the differential item functioning measures when $D=1$ under the two-parameter logistic model ($L=27$).

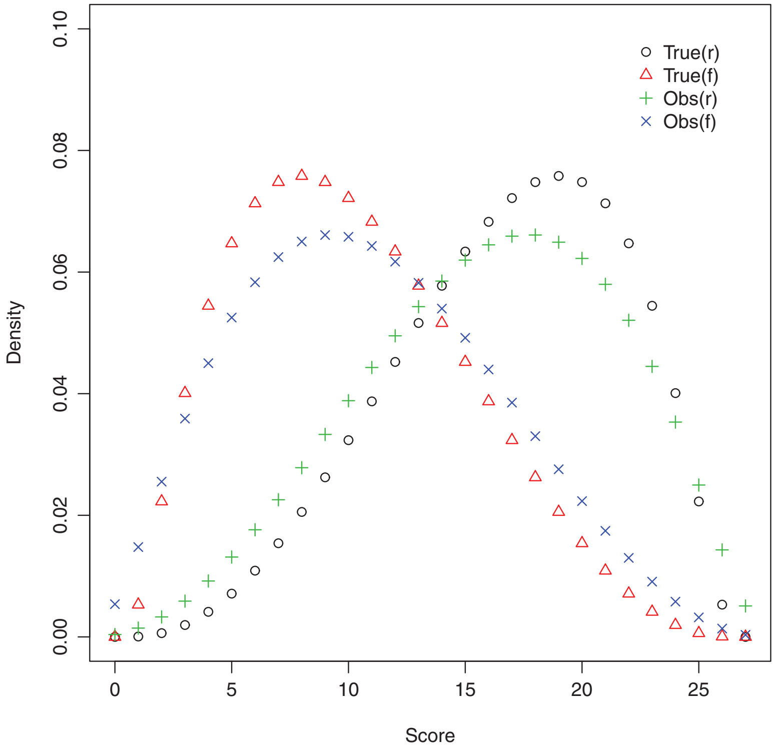

Figure 3 Density function of the true score and the observed score when $D=1$ under the two-parameter logistic model ($L=27$).

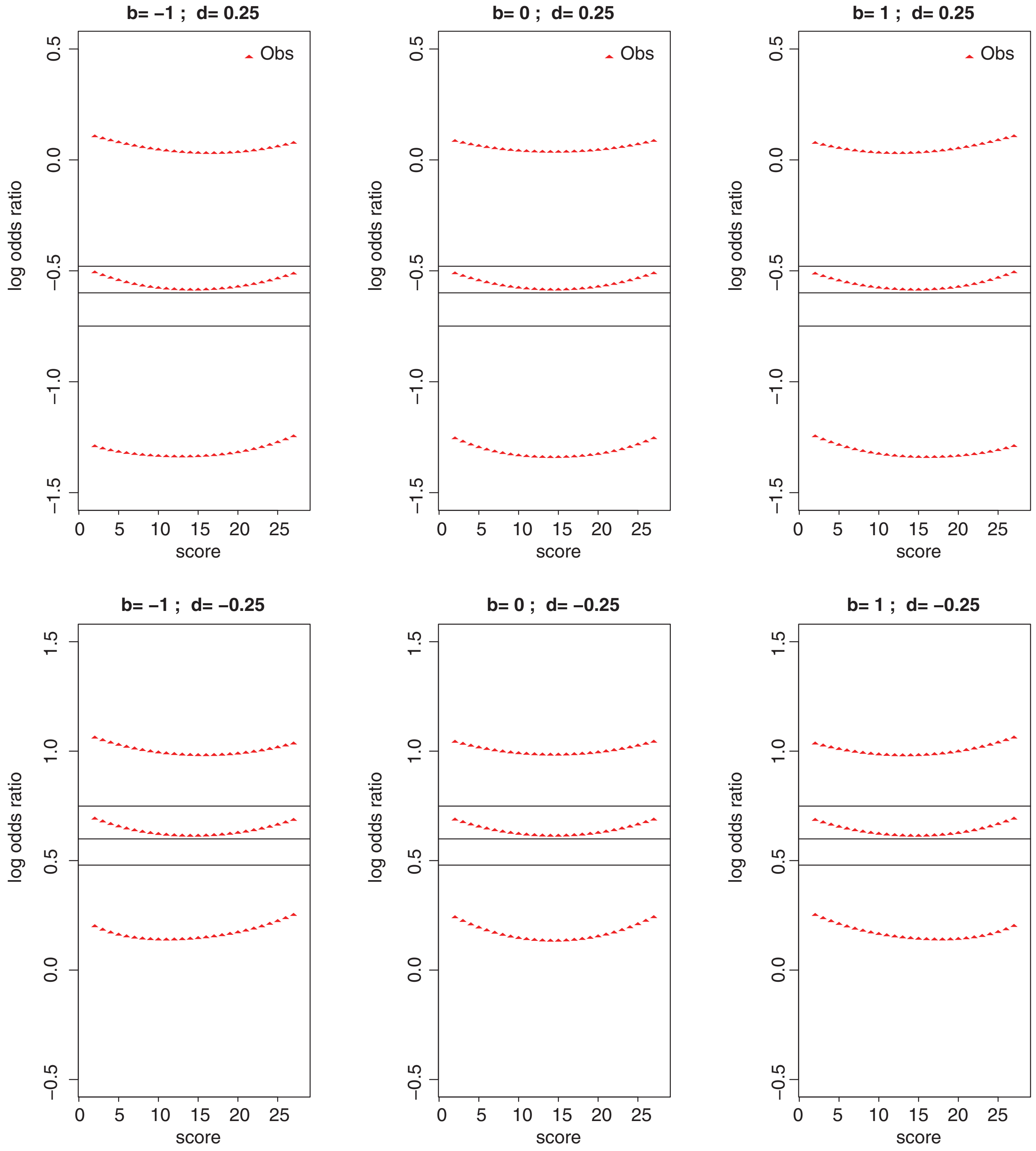

Figure 4 Conditional log odds ratios based on X when $D=1$ under the two-parameter logistic model for a short test ($L=27$). Panels show differential item functioning cases for $d=.25$ (upper panels) and $d=-.25$ (lower panels). Within each panel, the three curves made of triangles from top to bottom stand for items of $a=.48$, .60, and .75, respectively.

Table 4 Differential Item Functioning Measures When $D = 1$ Under the Two-Parameter Logistic Model ($L = 27$) With the Weighted Sum Scores

Item	a	b_r	d	$\Delta - \text{DIF}_\theta$	$\Delta - \text{DIF}_X$	$d(\Delta)$	$P - \text{DIF}_\theta$	$P - \text{DIF}_X$	$d(P)$
1	0.48	−1	0.25	−0.4794	−0.4616	0.0178	−0.0438	−0.0394	0.0044
2	0.48	−1	0.00	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
3	0.48	−1	−0.25	0.4794	0.4616	−0.0178	0.0425	0.0382	−0.0043
4	0.48	0	0.25	−0.4794	−0.4616	0.0178	−0.0425	−0.0387	0.0038
5	0.48	0	0.00	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
6	0.48	0	−0.25	0.4794	0.4616	−0.0178	0.0438	0.0398	−0.0040
7	0.48	1	0.25	−0.4794	−0.4616	0.0178	−0.0331	−0.0305	0.0026
8	0.48	1	0.00	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
9	0.48	1	−0.25	0.4794	0.4616	−0.0178	0.0359	0.0332	−0.0027
10	0.60	−1	0.25	−0.5992	−0.5644	0.0348	−0.0512	−0.0455	0.0057
11	0.60	−1	0.00	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
12	0.60	−1	−0.25	0.5993	0.5644	−0.0349	0.0493	0.0436	−0.0057
13	0.60	0	0.25	−0.5992	−0.5644	0.0348	−0.0493	−0.0446	0.0047
14	0.60	0	0.00	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
15	0.60	0	−0.25	0.5993	0.5644	−0.0349	0.0512	0.0463	−0.0049
16	0.60	1	0.25	−0.5992	−0.5644	0.0348	−0.0352	−0.0322	0.0030
17	0.60	1	0.00	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
18	0.60	1	−0.25	0.5993	0.5644	−0.0349	0.0393	0.0361	−0.0032
19	0.75	−1	0.25	−0.7491	−0.6851	0.0639	−0.0587	−0.0511	0.0076
20	0.75	−1	0.00	−0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
21	0.75	−1	−0.25	0.7491	0.6851	−0.0639	0.0559	0.0483	−0.0076
22	0.75	0	0.25	−0.7491	−0.6851	0.0639	−0.0559	−0.0500	0.0059
23	0.75	0	0.00	−0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
24	0.75	0	−0.25	0.7491	0.6851	−0.0639	0.0587	0.0526	−0.0062
25	0.75	1	0.25	−0.7491	−0.6851	0.0639	−0.0361	−0.0326	0.0035
26	0.75	1	0.00	−0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
27	0.75	1	−0.25	0.7491	0.6851	−0.0639	0.0417	0.0378	−0.0039

DIF exists, the magnitude of $d(\Delta) = \Delta - \text{DIF}_X - \Delta - \text{DIF}_\theta$ and $d(P) = P - \text{DIF}_X - P - \text{DIF}_\theta$ decreases with the test length but has a relatively large value, except for $a = .60$.

Two-Parameter Logistic Model With Weighted Sum Scores

In the 2PL case, the weighted sum score, $\sum_j a_j Y_j$, is a sufficient statistic for the latent ability (Lord, 1980, p. 57) for given item parameters. Therefore we expect that the weighted sum score-based and the true score-based DIF measures should be similar.

Tables 4, 5, and 6 for the test length of 27, 54, and 108, respectively, show the DIF measures when $D = \mu_f - \mu_r = 1$ and when the weighted sum scores are used as the matching variable.

As with the 1PL cases, all the DIF measures are zero under the null hypothesis. Under the alternative hypothesis, the observed score-based DIF measures are regressed more to zero compared to the true score-based DIF measures when DIF exists. Again, the overall DIF size increases in magnitude as a increases. In addition, $\Delta - \text{DIF}_X$ is not affected by item difficulty b_r or the DIF size $d = b_f - b_r$. However, the difference between $\Delta - \text{DIF}_X$ and $\Delta - \text{DIF}_\theta$ increases in magnitude as a increases, but it seems tolerable (to the second decimal place) given the large ability difference in the focal and reference groups. The overall difference between $P - \text{DIF}_X$ and $P - \text{DIF}_\theta$ increases as a increases, but it is small in magnitude as well (to the third decimal place). For given a, the P-DIF measures decrease in absolute value as the relative item difficulty $b^* = b_r + d - \mu_f$ increases in absolute value. As the test length increases, the differences between the observed score-based DIF measures and those based on true scores diminish.

Summary of Results

As seen in the section Analytical Results, differences between the numerical values of $\Delta - \text{DIF}_\theta$ and the theoretical values of $-4a(b_f - b_r)$ were mostly to the fourth decimal places, which were caused by the numerical approximation of

Table 5 Differential Item Functioning Measures When $D = 1$ Under the Two-Parameter Logistic Model ($L = 54$) With the Weighted Sum Scores

Item	a	b_r	d	$\Delta - \mathrm{DIF}_\theta$	$\Delta - \mathrm{DIF}_X$	$d(\Delta)$	$P - \mathrm{DIF}_\theta$	$P - \mathrm{DIF}_X$	$d(P)$
1	0.48	−1	0.25	−0.4794	−0.4767	0.0027	−0.0438	−0.0427	0.0011
2	0.48	−1	0.00	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
3	0.48	−1	−0.25	0.4794	0.4767	−0.0027	0.0425	0.0414	−0.0011
4	0.48	0	0.25	−0.4794	−0.4767	0.0027	−0.0425	−0.0416	0.0009
5	0.48	0	0.00	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
6	0.48	0	−0.25	0.4794	0.4767	−0.0027	0.0438	0.0428	−0.0009
7	0.48	1	0.25	−0.4794	−0.4767	0.0027	−0.0331	−0.0325	0.0006
8	0.48	1	0.00	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
9	0.48	1	−0.25	0.4794	0.4767	−0.0027	0.0359	0.0353	−0.0006
10	0.60	−1	0.25	−0.5992	−0.5918	0.0074	−0.0512	−0.0496	0.0017
11	0.60	−1	0.00	−0.0000	−0.0000	−0.0000	0.0000	−0.0000	−0.0000
12	0.60	−1	−0.25	0.5993	0.5918	−0.0074	0.0493	0.0476	−0.0017
13	0.60	0	0.25	−0.5992	−0.5918	0.0074	−0.0493	−0.0480	0.0013
14	0.60	0	0.00	−0.0000	−0.0000	−0.0000	0.0000	−0.0000	−0.0000
15	0.60	0	−0.25	0.5993	0.5918	−0.0074	0.0512	0.0498	−0.0014
16	0.60	1	0.25	−0.5992	−0.5918	0.0074	−0.0352	−0.0344	0.0008
17	0.60	1	0.00	−0.0000	−0.0000	−0.0000	0.0000	−0.0000	−0.0000
18	0.60	1	−0.25	0.5993	0.5918	−0.0074	0.0393	0.0384	−0.0009
19	0.75	−1	0.25	−0.7491	−0.7313	0.0178	−0.0587	−0.0562	0.0025
20	0.75	−1	0.00	−0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
21	0.75	−1	−0.25	0.7491	0.7313	−0.0178	0.0559	0.0533	−0.0025
22	0.75	0	0.25	−0.7491	−0.7313	0.0178	−0.0559	−0.0540	0.0019
23	0.75	0	0.00	−0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
24	0.75	0	−0.25	0.7491	0.7313	−0.0178	0.0587	0.0567	−0.0020
25	0.75	1	0.25	−0.7491	−0.7313	0.0178	−0.0361	−0.0350	0.0011
26	0.75	1	0.00	−0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
27	0.75	1	−0.25	0.7491	0.7313	−0.0178	0.0417	0.0404	−0.0012

integration in Equation 15; these differences should be independent of the test length and group difference by definition. However, the differences between $\Delta - \mathrm{DIF}_X$ and $\Delta - \mathrm{DIF}_\theta$, or between $P - \mathrm{DIF}_X$ and $P - \mathrm{DIF}_\theta$, depend on the test length. In the theoretical result of Equation 14, when X is the sufficient statistic for θ, impact of the test length is canceled out for $\Delta - \mathrm{DIF}_X$ by mathematical manipulations (refer to Appendix B); however, when computing Equation 4 analytically, such a cancellation does not occur. Unlike θ, a hypothetical latent variable, the manifest variable X is a function of the finite test length, among other factors; when the test is longer, X approaches the true score defined in Equation 11. Hence the differences between the observed score- and true score-based DIF measures are reduced when the test length is longer.

We summarize the results as follows.

Under the One-Parameter Logistic Model

As observed in Tables 1 and A1, both $\Delta - \mathrm{DIF}_X$ and $P - \mathrm{DIF}_X$ were zero under the null hypothesis; the observed score-based DIF was closer to zero than the true score-based DIF (because $X = T + \varepsilon$, and the discrete observed scores obscure the fine differences in the continuous variable θ):

- The differences of $\Delta - \mathrm{DIF}_X - \Delta - \mathrm{DIF}_\theta$ were the same and negligible regardless of the values of $D = \mu_r - \mu_f$, $d = b_f - b_r$, and b_r. This difference converged to zero as the test length increased; this was true for a test with the reliability of .95 in our design.
- The magnitude of $P - \mathrm{DIF}_X - P - \mathrm{DIF}_\theta$ was the same and negligible regardless of the value of D. This difference converged to zero as the test length increased. In addition, as the relative item difficulty to the focal group $b^* = b_r + d - \mu_f$ increased in absolute value, $P - \mathrm{DIF}_X$, $P - \mathrm{DIF}_\theta$, and $d(P) = P - \mathrm{DIF}_X - P - \mathrm{DIF}_\theta$ decreased in absolute value; this also led to the mirror effect on $d(P)$ between $D = 0$ and $D = 1$ in Tables 1 and A1.
- The convergence rate of $|\Delta - \mathrm{DIF}_X - \Delta - \mathrm{DIF}_\theta|$ seemed to be faster than that of $|P - \mathrm{DIF}_X - P - \mathrm{DIF}_\theta|$.

Table 6 Differential Item Functioning Measures When $D = 1$ Under the Two-Parameter Logistic Model ($L = 108$) With the Weighted Sum Scores

Item	a	b_r	d	$\Delta - \text{DIF}_\theta$	$\Delta - \text{DIF}_X$	$d(\Delta)$	$P - \text{DIF}_\theta$	$P - \text{DIF}_X$	$d(P)$
1	0.48	−1	0.25	−0.4794	−0.4791	0.0003	−0.0438	−0.0434	0.0003
2	0.48	−1	0.00	−0.0000	−0.0000	−0.0000	0.0000	−0.0000	−0.0000
3	0.48	−1	−0.25	0.4794	0.4791	−0.0003	0.0425	0.0422	−0.0003
4	0.48	0	0.25	−0.4794	−0.4791	0.0003	−0.0425	−0.0422	0.0003
5	0.48	0	0.00	−0.0000	−0.0000	−0.0000	0.0000	−0.0000	−0.0000
6	0.48	0	−0.25	0.4794	0.4791	−0.0003	0.0438	0.0434	−0.0003
7	0.48	1	0.25	−0.4794	−0.4791	0.0003	−0.0331	−0.0329	0.0002
8	0.48	1	0.00	−0.0000	−0.0000	−0.0000	0.0000	−0.0000	−0.0000
9	0.48	1	−0.25	0.4794	0.4791	−0.0003	0.0359	0.0357	−0.0002
10	0.60	−1	0.25	−0.5992	−0.5982	0.0011	−0.0512	−0.0506	0.0006
11	0.60	−1	0.00	−0.0000	−0.0000	−0.0000	0.0000	−0.0000	−0.0000
12	0.60	−1	−0.25	0.5993	0.5982	−0.0011	0.0493	0.0487	−0.0006
13	0.60	0	0.25	−0.5992	−0.5982	0.0011	−0.0493	−0.0488	0.0005
14	0.60	0	0.00	−0.0000	−0.0000	−0.0000	0.0000	−0.0000	−0.0000
15	0.60	0	−0.25	0.5993	0.5982	−0.0011	0.0512	0.0507	−0.0005
16	0.60	1	0.25	−0.5992	−0.5982	0.0011	−0.0352	−0.0349	0.0003
17	0.60	1	0.00	−0.0000	−0.0000	−0.0000	0.0000	−0.0000	−0.0000
18	0.60	1	−0.25	0.5993	0.5982	−0.0011	0.0393	0.0390	−0.0003
19	0.75	−1	0.25	−0.7491	−0.7457	0.0034	−0.0587	−0.0578	0.0010
20	0.75	−1	0.00	−0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
21	0.75	−1	−0.25	0.7491	0.7457	−0.0034	0.0559	0.0549	−0.0010
22	0.75	0	0.25	−0.7491	−0.7457	0.0034	−0.0559	−0.0551	0.0008
23	0.75	0	0.00	−0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
24	0.75	0	−0.25	0.7491	0.7457	−0.0034	0.0587	0.0579	−0.0009
25	0.75	1	0.25	−0.7491	−0.7457	0.0034	−0.0361	−0.0356	0.0005
26	0.75	1	0.00	−0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
27	0.75	1	−0.25	0.7491	0.7457	−0.0034	0.0417	0.0412	−0.0005

Under the Two-Parameter Logistic Model and Matched on the Simple (Unweighted) Sum Scores

When $\mu_f - \mu_r = 0$, both $\Delta - \text{DIF}_X$ and $P - \text{DIF}_X$ were zero under the null hypothesis. Under the alternate hypothesis, the DIF measure $\Delta - \text{DIF}_X$ was closer to zero than $\Delta - \text{DIF}_\theta$; all the DIF measures increased (in absolute value) as a increased:

- The values of $\Delta - \text{DIF}_X - \Delta - \text{DIF}_\theta$ were small (to the third decimal place), and the smallest occurred when $a = .60$, the average discrimination. The difference became smaller as the test length increased, but it did not converge to zero when $a \neq .60$ for any test with a reasonably long length, as observed in Tables 2, A2, and A3. Under the 2PL model, for a test with a finite number of items, the odds ratios based on the simple sum scores varied across different values of X (i.e., DIF is not uniform). However, the differences between $\Delta - \text{DIF}_X$ and $\Delta - \text{DIF}_\theta$ were negligible.
- The values of $P - \text{DIF}_X - P - \text{DIF}_\theta$ were small (to the third decimal place), and they became smaller as the test length increased. The differences, however, did not converge to zero for any test with a reasonably long length, as observed in Tables 2, A2, and A3. This effect is caused by the fact that X is not a sufficient statistic for θ; that is, $P_r(Y = 1 | X)$ is a function of θ, and $P_f(Y = 1 | X)$ is a function of $\theta + d$. However, the differences seemed negligible. In addition, for a fixed a parameter, as the relative item difficulty $b^* = b_r + d - \mu_f$ increased in absolute value, $|P - \text{DIF}_X - P - \text{DIF}_\theta|$ decreased.

When $\mu_f - \mu_r \neq 0$, the observed score-based DIF measures and the true score-based DIF measures were substantially different; the difference was positive when $a = .48$ and negative when $a = .75$:

- The values of $\Delta - \text{DIF}_X - \Delta - \text{DIF}_\theta$ were relatively large (to the first decimal place), except for $a = .60$, where the difference was to the second decimal place. The difference became smaller as the test length increased, but it did not converge to zero for any test with a reasonably long length, as observed in Tables 3, A4, and A5, because under the 2PL model, the odds ratios based on observed scores varied for different X and substantially deviated from the analytical values.

- The values of $P-\text{DIF}_X-P-\text{DIF}_\theta$ were relatively large (to the second decimal place), except for when $a=.60$, where the difference was to the third decimal place. The difference became smaller as the test length increased, but it did not converge to zero for any test with a reasonably long length, as observed in Tables 3, A4, and A5, because X is not a sufficient statistic for θ. The conditional probability $P_r(Y=1|X)$ is a function of θ, and $P_f(Y=1|X)$ is a function of $\theta-D+d$.
- Even under the null hypothesis, when $D=1$, both $\Delta-\text{DIF}_X$ and $P-\text{DIF}_X$ may not be zero for any practical test with a finite number of items.

Under the Two-Parameter Logistic Model and Matching on the Weighted Sum Scores

As observed in Tables 4, 5, and 6, when matching on the weighted sum scores, results under the 2PL model for $D=1$ were somewhat similar to cases under the 1PL model. Both $\Delta-\text{DIF}_X$ and $P-\text{DIF}_X$ were zero under the null hypothesis. The observed score-based DIF was closer to zero than the true score-based DIF under the alternative hypothesis:

- The values of $\Delta-\text{DIF}_X-\Delta-\text{DIF}_\theta$ increased in absolute value as a increased but remained at a tolerable magnitude (to the second decimal place). The difference seemed as if it would converge to zero as the test length increased, as suggested by Tables 4, 5, and 6.
- The magnitude of $P-\text{DIF}_X-P-\text{DIF}_\theta$ increased in absolute value as a increased but still remained at a tolerable magnitude (to the third decimal place). The difference appeared to converge to zero as the test length increased. In addition, as the relative item difficulty $|b_r+d-\mu_f|$ increased, $|P-\text{DIF}_X-P-\text{DIF}_\theta|$ decreased.

Discussion

In previous DIF studies of the MH D-DIF and STD P-DIF statistics, the simple sum score has often been used as the matching variable, and $\Delta-\text{DIF}_\theta$ and $P-\text{DIF}_\theta$ have been used as criteria to evaluate performance of these DIF statistics. Most of those studies simulated response data from IRT models and found biases. In this study, we analytically derived formulas for DIF measures $\Delta-\text{DIF}_X$ and $P-\text{DIF}_X$ for general item response models. The differences between observed score- and true score-based DIF measures were presented for the 1PL and 2PL IRT models. Two weighting schemes were studied to construct the matching variable: the simple sum score (unweighted or the uniformly weighted sum score) and the sum score weighted by item discrimination parameters. Because the results are based on analytical formulas, they can be used to investigate different DIF conditions beyond the cases reported in this study under more general item response models when the item response function $P(Y=1|\theta)$ is available.

When the simple sum score is the matching variable, our illustrative examples demonstrated that, given adequate test length, the practice of using $\Delta-\text{DIF}_\theta$ and $P-\text{DIF}_\theta$ as criteria is acceptable when the data follow a 1PL model. When data follow the 2PL model and DIF exists, the DIF measure $\Delta-\text{DIF}_X$ is always different from $\Delta-\text{DIF}_\theta$ for any test with a finite number of items, particularly for items with very different discrimination from the average test discrimination. Nevertheless, these differences may not be a cause for practical concern when the ability difference between the focal group and the reference group is small.

However, when the groups are different in ability and the data follow the 2PL model, which is a common condition in simulation studies, these simple sum score-based DIF measures may not be zero even under the null hypothesis of no DIF for any realistic tests with finite numbers of items. The magnitude of the bias may have practical impact, particularly for items that deviate from the average discrimination of the test.

Based on these results, we conclude that the most important influence on the simple sum score-based DIF measures, under the 2PL model, is the ability difference between the focal and reference groups, which leads to the observation that $\Delta-\text{DIF}_X\neq 0$ and $P-\text{DIF}_X\neq 0$ under the null hypothesis of no DIF. The next important influence is variation of the item discrimination parameter a. When the studied item has a discrimination parameter close to the average discrimination of the test, the simple sum score-based DIF measures and the true score-based DIF measures exhibit small differences. The relative item difficulty to the focal group may play a smaller role: When the item is relatively harder, the differences between the observed score-based DIF measures and the true score-based DIF measures are slightly smaller.

Our results show analytically that matching on sufficient statistics for the latent true ability should in principle improve the performance of DIF statistics, leading to the two weighted sum score-based DIF measures converging to the true score-based DIF criteria at the test length increases. Even though numerical computation of DIF measures with weighted sum scores is quite challenging, they are straightforward when item response data and item discrimination parameters are available.

One limitation of our study is related to the general iterative formula for computing the exact distribution of the weighted sum score. When the item discrimination parameters are all different, computation of the exact distribution could be very time consuming (in this case, matching on the exact weighted score is similar to matching on item response patterns). Further investigation on fast algorithms is valuable to approximate such a distribution, for instance, by placing weighted sums into bins.

Several practical issues are associated with using the weighted sum scores as a matching variable. The effect of substituting estimates for parameters needs investigation. Sample size is obviously important here, and model fit is also likely to affect the results. In addition, if the 2PL does not hold or the sample size is too small to get good estimates of the discrimination parameters, using the weighted sum score may be worse than using the simple sum.

The biggest challenge to using our analytical results in practice is how to achieve exact matching as the number of possible values that the weighted sum can assume becomes very large. Even if exact matching were achieved with a large sample of test takers, the results could prove unstable. In future research, we will investigate the effects of replacing exact matching with various coarse matches of the weighted sum score on the fidelity of the DIF results to that obtained with exact matching. Before applying the weighted sum as a matching variable in DIF practice, a number of areas clearly need to be studied.

Notes

[1] One can replace $g_r(x)$ by $g_T(x)$, the score distribution of the total group, in Equation 4. However, the change in $\Delta - \text{DIF}_X$ is less than .007 in the worst scenario (refer to Table 3). The replacement of $g_r(\theta)$ by $g_T(\theta)$ in Equation 3 does not have an impact because of a constant ratio with respect to θ. Therefore we kept the Zwick et al. formula in this study.

[2] Note that whereas the MH D-DIF statistic represents an arithmetic average, $\Delta - \text{DIF}_\theta$ and $\Delta - \text{DIF}_X$ are geometric averages. Our study does not address this discrepancy. Instead, we compare Equations 3 and 4.

[3] Using 81 nodes on a larger interval increased the accuracy by approximately .001 for the case studied in the section When $D = \mu_r - \mu_f = 1$.

References

Agresti, A. (2002). *Categorical data analysis*. New York, NY: John Wiley.

Allen, N. L., & Donoghue, J. R. (1996). Applying the Mantel–Haenszel procedure to complex samples of items. *Journal of Educational Measurement, 33*, 231–251. https://doi.org/10.1111/j.1745-3984.1996.tb00491.x

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

Chang, H. H., Mazzeo, J., & Roussos, L. (1995). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *Journal of Educational Measurement, 33*, 333–353. https://doi.org/10.1002/j.2333-8504.1995.tb01640.x

Cressie, N., & Holland, P. (1981). Characterizing the manifest probabilities of latent trait models. *Psychometrika, 48*, 129–141. https://doi.org/10.1007/BF02314681

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). Hillsdale, NJ: Erlbaum.

Dorans, N. (2013). *ETS contributions to the quantitative assessment of item, test, and score fairness* (Research Report No. RR-13-27). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2013.tb02334.x

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel–Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Hillsdale, NJ: Erlbaum.

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

Haberman, S. (2013). *A general program for item-response analysis that employs the stabilized Newton–Raphson algorithm* (Research Report No. RR-13-32). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2013.tb02339.x

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

Holland, P., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Erlbaum.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, linking, and scaling: Methods and practices* (3rd ed.). New York, NY: Springer.

Kolen, M. J., Zeng, L., & Hanson, B. A. (1996). Conditional CSEM of the scale scores for scale scores using IRT. *Journal of Educational Measurement, 33*, 129 – 140. https://doi.org/10.1080/15305058.2011.617476

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

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed score equatings. *Applied Psychological Measurement, 8*, 453 – 461. https://doi.org/10.1177/014662168400800409

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

Meredith, W., & Millsap, R. (1992). On the misuse of manifest variables in the detection of measurement bias. *Psychometrika, 57*, 289 – 311. https://doi.org/10.1007/BF02294510

Penfield, R. (2007). An approach for categorizing DIF in polytomous items. *Applied Measurement in Education, 20*, 335 – 355. https://doi.org/10.1080/08957340701431435

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

Roussos, L. A., & Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel – Haenszel type 1 error performance. *Journal of Educational Measurement, 33*, 215 – 230. https://doi.org/10.1111/j.1745-3984.1996.tb00490.x

Shealy, R., & Stout, W. F. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58*, 159 – 194. https://doi.org/10.1007/BF02294572

Spray, J., & Miller, T. (1992). *Performance of the Mantel – Haenszel statistic and the standardized difference in proportion correct when population ability distributions are incongruent* (Research Report No. 92-1). Iowa City, IA: American College Testing.

Thissen, D., Pommerich, M., Billeaud, K., & Willams, V. (1995). Item response theory for scores on test including polytomous items with ordered responses. *Applied Psychological Measurement, 19*, 39 – 49. https://doi.org/10.1177/014662169501900105

Verhelst, N. D., & Glas, C. A. W. (1995). The one parameter logistic model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Their foundations, recent developments and applications* (pp. 215 – 237). New York, NY: Springer.

Zieky, M. (1993). DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337 – 347). Hillsdale, NJ: Erlbaum.

Zieky, M. (2011). The origins of procedures for using differential item functioning statistics at Educational Testing Service. In N. J. Dorans & S. Sinharay (Eds.), *Looking back: Proceedings of a conference in honor of Paul W. Holland* (pp. 115 – 127). New York, NY: Springer.

Zwick, R. (1990). When do item response function and Mantel – Haensel definitions of differential item functioning coincide? *Journal of Educational Statistics, 15*, 185 – 197. https://doi.org/10.3102/10769986015003185

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). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2012.tb02290.x

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

Zwick, R., Thayer, D. T., & Mazzeo, J. (1997). *Describing and categorizing DIF in polytomous items* (Research Report No. RR-97-05). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1997.tb01726.x

Appendix A

Relevant Figures and Tables

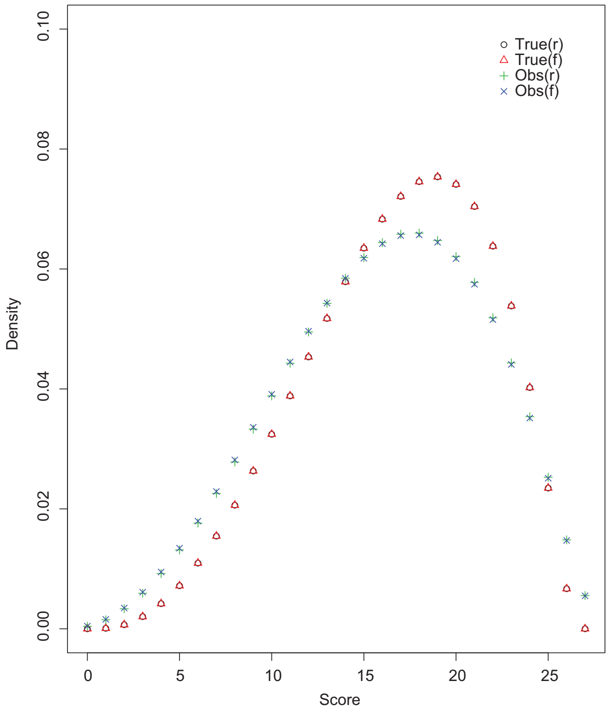

Figure A1 Density functions of the true and observed scores for the focal and reference groups for a short test ($L = 27$) under the one-parameter logistic model when the group ability difference $D = 1$. Because of the short test, differences exist between the observed and true scores.

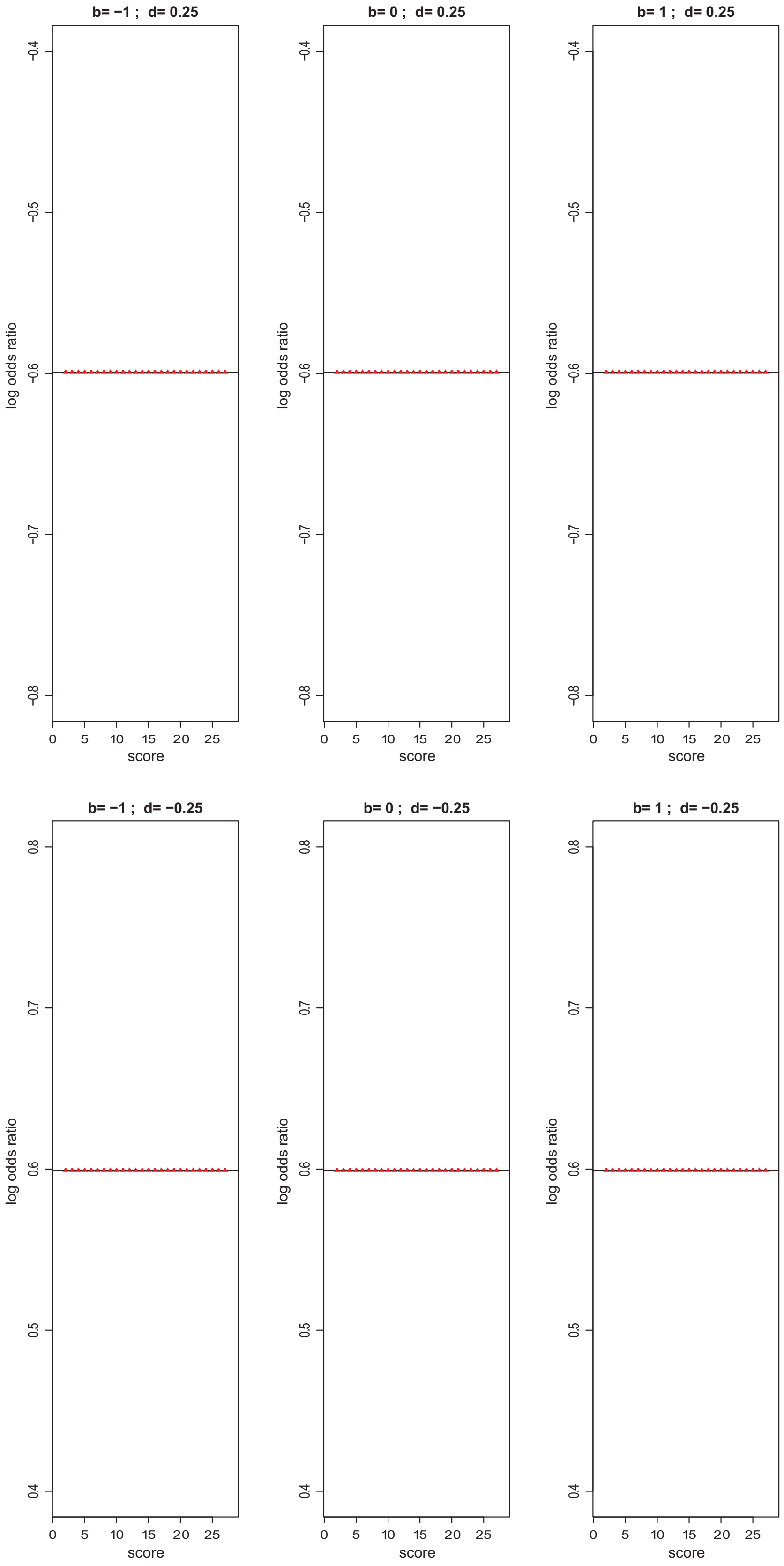

Figure A2 Conditional log odds ratio based on X when $D = 1$ under the one-parameter logistic model for a short test ($L = 27$). The odds ratio is constant across X.

Table A1 Differential Item Functioning Measures Under the One-Parameter Logistic Model for a Medium ($L = 54$, Reliability .91) and a Long ($L = 108$, Reliability .95) Test

Item	a	b_r	d	$\Delta - \mathrm{DIF}_\theta$	$\Delta - \mathrm{DIF}_X$	$d(\Delta)$	$P - \mathrm{DIF}_\theta$	$P - \mathrm{DIF}_X$	$d(P)$
$L = 54, D = 0$									
1	0.60	−1	0.25	−0.5992	−0.5987	0.0005	−0.0393	−0.0386	0.0007
2	0.60	−1	0.00	−0.0000	−0.0000	0.0000	0.0000	0.0000	0.0000
3	0.60	−1	−0.25	0.5993	0.5987	−0.0005	0.0352	0.0346	−0.0006
4	0.60	0	0.25	−0.5992	−0.5987	0.0005	−0.0512	−0.0502	0.0010
5	0.60	0	0.00	−0.0000	−0.0000	0.0000	0.0000	0.0000	0.0000
6	0.60	0	−0.25	0.5993	0.5987	−0.0005	0.0493	0.0483	−0.0010
7	0.60	1	0.25	−0.5992	−0.5987	0.0005	−0.0493	−0.0482	0.0011
8	0.60	1	0.00	−0.0000	−0.0000	0.0000	0.0000	0.0000	0.0000
9	0.60	1	−0.25	0.5993	0.5987	−0.0005	0.0512	0.0501	−0.0011
$L = 54, D = 1$									
1	0.60	−1	0.25	−0.5992	−0.5987	0.0005	−0.0512	−0.0501	0.0011
2	0.60	−1	0.00	−0.0000	−0.0000	−0.0000	0.0000	−0.0000	−0.0000
3	0.60	−1	−0.25	0.5993	0.5987	−0.0005	0.0493	0.0482	−0.0011
4	0.60	0	0.25	−0.5992	−0.5987	0.0005	−0.0493	−0.0483	0.0010
5	0.60	0	0.00	−0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
6	0.60	0	−0.25	0.5993	0.5987	−0.0005	0.0512	0.0502	−0.0010
7	0.60	1	0.25	−0.5992	−0.5987	0.0005	−0.0352	−0.0346	0.0006
8	0.60	1	0.00	−0.0000	−0.0000	−0.0000	0.0000	0.0000	0.0000
9	0.60	1	−0.25	0.5993	0.5987	−0.0005	0.0393	0.0386	−0.0007
$L = 108, D = 0$									
1	0.60	−1	0.25	−0.5992	−0.5992	0.0000	−0.0393	−0.0390	0.0003
2	0.60	−1	0.00	−0.0000	−0.0000	0.0000	0.0000	0.0000	0.0000
3	0.60	−1	−0.25	0.5993	0.5992	−0.0000	0.0352	0.0349	−0.0003
4	0.60	0	0.25	−0.5992	−0.5992	0.0000	−0.0512	−0.0507	0.0005
5	0.60	0	0.00	−0.0000	−0.0000	0.0000	0.0000	0.0000	0.0000
6	0.60	0	−0.25	0.5993	0.5992	−0.0000	0.0493	0.0488	−0.0005
7	0.60	1	0.25	−0.5992	−0.5992	0.0000	−0.0493	−0.0487	0.0006
8	0.60	1	0.00	−0.0000	−0.0000	0.0000	0.0000	0.0000	0.0000
9	0.60	1	−0.25	0.5993	0.5992	−0.0000	0.0512	0.0506	−0.0006
$L = 108, D = 1$									
1	0.60	−1	0.25	−0.5992	−0.5992	0.0000	−0.0512	−0.0506	0.0006
2	0.60	−1	0.00	−0.0000	−0.0000	−0.0000	0.0000	−0.0000	−0.0000
3	0.60	−1	−0.25	0.5993	0.5992	−0.0000	0.0493	0.0487	−0.0006
4	0.60	0	0.25	−0.5992	−0.5992	0.0000	−0.0493	−0.0488	0.0005
5	0.60	0	0.00	−0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
6	0.60	0	−0.25	0.5993	0.5992	−0.0000	0.0512	0.0507	−0.0005
7	0.60	1	0.25	−0.5992	−0.5992	0.0000	−0.0352	−0.0349	0.0003
8	0.60	1	0.00	−0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
9	0.60	1	−0.25	0.5993	0.5992	−0.0000	0.0393	0.0390	−0.0003

Table A2 Differential Item Functioning Measures When $D = 0$ Under the Two-Parameter Logistic Model for a Test of Length $L = 54$

Item	a	b_r	d	$\Delta - \mathrm{DIF}_\theta$	$\Delta - \mathrm{DIF}_X$	$d(\Delta)$	$P - \mathrm{DIF}_\theta$	$P - \mathrm{DIF}_X$	$d(P)$
1	0.48	−1	0.25	−0.4794	−0.4753	0.0041	−0.0359	−0.0340	0.0020
2	0.48	−1	0.00	−0.0000	−0.0000	0.0000	0.0000	0.0000	0.0000
3	0.48	−1	−0.25	0.4794	0.4756	−0.0038	0.0331	0.0316	−0.0015
4	0.48	0	0.25	−0.4794	−0.4747	0.0047	−0.0438	−0.0401	0.0037
5	0.48	0	0.00	−0.0000	−0.0000	0.0000	0.0000	0.0000	0.0000
6	0.48	0	−0.25	0.4794	0.4748	−0.0046	0.0425	0.0391	−0.0034
7	0.48	1	0.25	−0.4794	−0.4748	0.0046	−0.0425	−0.0391	0.0034
8	0.48	1	0.00	−0.0000	−0.0000	0.0000	0.0000	0.0000	0.0000
9	0.48	1	−0.25	0.4794	0.4747	−0.0047	0.0438	0.0401	−0.0037
10	0.60	−1	0.25	−0.5992	−0.5988	0.0005	−0.0393	−0.0386	0.0007
11	0.60	−1	0.00	−0.0000	−0.0000	0.0000	0.0000	0.0000	0.0000
12	0.60	−1	−0.25	0.5993	0.5988	−0.0005	0.0352	0.0345	−0.0007
13	0.60	0	0.25	−0.5992	−0.5988	0.0005	−0.0512	−0.0500	0.0012
14	0.60	0	0.00	−0.0000	−0.0000	0.0000	0.0000	0.0000	0.0000
15	0.60	0	−0.25	0.5993	0.5988	−0.0005	0.0493	0.0481	−0.0011
16	0.60	1	0.25	−0.5992	−0.5988	0.0005	−0.0493	−0.0481	0.0012
17	0.60	1	0.00	−0.0000	−0.0000	0.0000	0.0000	0.0000	0.0000
18	0.60	1	−0.25	0.5993	0.5988	−0.0005	0.0512	0.0500	−0.0012
19	0.75	−1	0.25	−0.7491	−0.7426	0.0065	−0.0417	−0.0413	0.0003
20	0.75	−1	0.00	−0.0000	−0.0000	0.0000	0.0000	0.0000	0.0000
21	0.75	−1	−0.25	0.7491	0.7434	−0.0056	0.0361	0.0351	−0.0010
22	0.75	0	0.25	−0.7491	−0.7396	0.0095	−0.0587	−0.0613	−0.0025
23	0.75	0	0.00	−0.0000	−0.0000	0.0000	0.0000	0.0000	0.0000
24	0.75	0	−0.25	0.7491	0.7401	−0.0090	0.0559	0.0578	0.0019
25	0.75	1	0.25	−0.7491	−0.7400	0.0091	−0.0559	−0.0577	−0.0018
26	0.75	1	0.00	−0.0000	−0.0000	0.0000	0.0000	0.0000	0.0000
27	0.75	1	−0.25	0.7491	0.7395	−0.0096	0.0587	0.0612	0.0024

Table A3 Differential Item Functioning Measures When $D = 0$ Under the Two-Parameter Logistic Model for a Long Test ($L = 108$)

Item	a	b_r	d	$\Delta - \mathrm{DIF}_\theta$	$\Delta - \mathrm{DIF}_X$	$d(\Delta)$	$P - \mathrm{DIF}_\theta$	$P - \mathrm{DIF}_X$	$d(P)$
1	0.48	−1	0.25	−0.4794	−0.4757	0.0037	−0.0359	−0.0343	0.0016
2	0.48	−1	0.00	−0.0000	−0.0000	0.0000	0.0000	0.0000	0.0000
3	0.48	−1	−0.25	0.4794	0.4759	−0.0035	0.0331	0.0320	−0.0012
4	0.48	0	0.25	−0.4794	−0.4751	0.0043	−0.0438	−0.0405	0.0032
5	0.48	0	0.00	−0.0000	−0.0000	0.0000	0.0000	0.0000	0.0000
6	0.48	0	−0.25	0.4794	0.4752	−0.0042	0.0425	0.0395	−0.0030
7	0.48	1	0.25	−0.4794	−0.4751	0.0043	−0.0425	−0.0395	0.0030
8	0.48	1	0.00	−0.0000	−0.0000	0.0000	0.0000	0.0000	0.0000
9	0.48	1	−0.25	0.4794	0.4751	−0.0043	0.0438	0.0405	−0.0032
10	0.60	−1	0.25	−0.5992	−0.5992	0.0000	−0.0393	−0.0389	0.0004
11	0.60	−1	0.00	−0.0000	−0.0000	0.0000	0.0000	0.0000	0.0000
12	0.60	−1	−0.25	0.5993	0.5992	−0.0000	0.0352	0.0349	−0.0003
13	0.60	0	0.25	−0.5992	−0.5992	0.0000	−0.0512	−0.0505	0.0007
14	0.60	0	0.00	−0.0000	−0.0000	0.0000	0.0000	0.0000	0.0000
15	0.60	0	−0.25	0.5993	0.5992	−0.0000	0.0493	0.0486	−0.0006
16	0.60	1	0.25	−0.5992	−0.5992	0.0000	−0.0493	−0.0486	0.0006
17	0.60	1	0.00	−0.0000	−0.0000	0.0000	0.0000	0.0000	0.0000
18	0.60	1	−0.25	0.5993	0.5992	−0.0000	0.0512	0.0505	−0.0007
19	0.75	−1	0.25	−0.7491	−0.7431	0.0059	−0.0417	−0.0417	−0.0000
20	0.75	−1	0.00	−0.0000	−0.0000	0.0000	0.0000	0.0000	0.0000
21	0.75	−1	−0.25	0.7491	0.7440	−0.0051	0.0361	0.0354	−0.0007
22	0.75	0	0.25	−0.7491	−0.7402	0.0089	−0.0587	−0.0619	−0.0032
23	0.75	0	0.00	−0.0000	−0.0000	0.0000	0.0000	0.0000	0.0000
24	0.75	0	−0.25	0.7491	0.7407	−0.0084	0.0559	0.0584	0.0025
25	0.75	1	0.25	−0.7491	−0.7406	0.0085	−0.0559	−0.0584	−0.0025
26	0.75	1	0.00	−0.0000	−0.0000	0.0000	0.0000	0.0000	0.0000
27	0.75	1	−0.25	0.7491	0.7401	−0.0089	0.0587	0.0619	0.0031

Table A4 Differential Item Functioning Measures When $D = 1$ Under the Two-Parameter Logistic Model for a Test of Length $L = 54$

Item	a	b_r	d	$\Delta - \text{DIF}_\theta$	$\Delta - \text{DIF}_X$	$d(\Delta)$	$P - \text{DIF}_\theta$	$P - \text{DIF}_X$	$d(P)$
1	0.48	−1	0.25	−0.4794	0.0337	0.5131	−0.0438	0.0031	0.0468
2	0.48	−1	0.00	−0.0000	0.5093	0.5093	0.0000	0.0431	0.0431
3	0.48	−1	−0.25	0.4794	0.9851	0.5057	0.0425	0.0822	0.0397
4	0.48	0	0.25	−0.4794	0.0362	0.5156	−0.0425	0.0027	0.0452
5	0.48	0	0.00	−0.0000	0.5108	0.5108	0.0000	0.0418	0.0418
6	0.48	0	−0.25	0.4794	0.9856	0.5062	0.0438	0.0819	0.0382
7	0.48	1	0.25	−0.4794	0.0377	0.5171	−0.0331	0.0020	0.0351
8	0.48	1	0.00	−0.0000	0.5122	0.5122	0.0000	0.0337	0.0337
9	0.48	1	−0.25	0.4794	0.9866	0.5072	0.0359	0.0677	0.0318
10	0.60	−1	0.25	−0.5992	−0.5695	0.0298	−0.0512	−0.0477	0.0036
11	0.60	−1	0.00	−0.0000	0.0293	0.0293	0.0000	0.0023	0.0023
12	0.60	−1	−0.25	0.5993	0.6281	0.0289	0.0493	0.0504	0.0011
13	0.60	0	0.25	−0.5992	−0.5693	0.0299	−0.0493	−0.0461	0.0032
14	0.60	0	0.00	−0.0000	0.0294	0.0294	0.0000	0.0020	0.0020
15	0.60	0	−0.25	0.5993	0.6282	0.0290	0.0512	0.0521	0.0009
16	0.60	1	0.25	−0.5992	−0.5693	0.0299	−0.0352	−0.0331	0.0021
17	0.60	1	0.00	−0.0000	0.0294	0.0294	0.0000	0.0015	0.0015
18	0.60	1	−0.25	0.5993	0.6282	0.0290	0.0393	0.0400	0.0007
19	0.75	−1	0.25	−0.7491	−1.3109	−0.5619	−0.0587	−0.1065	−0.0478
20	0.75	−1	0.00	−0.0000	−0.5690	−0.5690	0.0000	−0.0453	−0.0453
21	0.75	−1	−0.25	0.7491	0.1737	−0.5753	0.0559	0.0124	−0.0435
22	0.75	0	0.25	−0.7491	−1.3110	−0.5619	−0.0559	−0.1053	−0.0494
23	0.75	0	0.00	−0.0000	−0.5713	−0.5713	0.0000	−0.0475	−0.0475
24	0.75	0	−0.25	0.7491	0.1687	−0.5804	0.0587	0.0137	−0.0451
25	0.75	1	0.25	−0.7491	−1.3150	−0.5659	−0.0361	−0.0666	−0.0305
26	0.75	1	0.00	−0.0000	−0.5739	−0.5739	0.0000	−0.0316	−0.0316
27	0.75	1	−0.25	0.7491	0.1664	−0.5826	0.0417	0.0097	−0.0320

Table A5 Differential Item Functioning Measures When $D = 1$ Under the Two-Parameter Logistic Model for a Long Test ($L = 108$)

Item	a	b_r	d	$\Delta - \text{DIF}_\theta$	$\Delta - \text{DIF}_X$	$d(\Delta)$	$P - \text{DIF}_\theta$	$P - \text{DIF}_X$	$d(P)$
1	0.48	−1	0.25	−0.4794	0.0294	0.5088	−0.0438	0.0026	0.0464
2	0.48	−1	0.00	−0.0000	0.5053	0.5053	0.0000	0.0431	0.0431
3	0.48	−1	−0.25	0.4794	0.9814	0.5020	0.0425	0.0827	0.0401
4	0.48	0	0.25	−0.4794	0.0307	0.5101	−0.0425	0.0023	0.0448
5	0.48	0	0.00	−0.0000	0.5058	0.5058	0.0000	0.0418	0.0418
6	0.48	0	−0.25	0.4794	0.9810	0.5016	0.0438	0.0823	0.0386
7	0.48	1	0.25	−0.4794	0.0319	0.5113	−0.0331	0.0017	0.0348
8	0.48	1	0.00	−0.0000	0.5067	0.5067	0.0000	0.0337	0.0337
9	0.48	1	−0.25	0.4794	0.9815	0.5021	0.0359	0.0680	0.0321
10	0.60	−1	0.25	−0.5992	−0.5706	0.0287	−0.0512	−0.0482	0.0030
11	0.60	−1	0.00	−0.0000	0.0286	0.0286	0.0000	0.0023	0.0023
12	0.60	−1	−0.25	0.5993	0.6278	0.0286	0.0493	0.0509	0.0017
13	0.60	0	0.25	−0.5992	−0.5705	0.0288	−0.0493	−0.0466	0.0027
14	0.60	0	0.00	−0.0000	0.0287	0.0287	0.0000	0.0021	0.0021
15	0.60	0	−0.25	0.5993	0.6279	0.0287	0.0512	0.0526	0.0014
16	0.60	1	0.25	−0.5992	−0.5704	0.0288	−0.0352	−0.0334	0.0018
17	0.60	1	0.00	−0.0000	0.0288	0.0288	0.0000	0.0015	0.0015
18	0.60	1	−0.25	0.5993	0.6279	0.0287	0.0393	0.0404	0.0011
19	0.75	−1	0.25	−0.7491	−1.3088	−0.5597	−0.0587	−0.1072	−0.0485
20	0.75	−1	0.00	−0.0000	−0.5663	−0.5663	0.0000	−0.0453	−0.0453
21	0.75	−1	−0.25	0.7491	0.1770	−0.5721	0.0559	0.0131	−0.0428
22	0.75	0	0.25	−0.7491	−1.3069	−0.5578	−0.0559	−0.1059	−0.0500
23	0.75	0	0.00	−0.0000	−0.5666	−0.5666	0.0000	−0.0475	−0.0475
24	0.75	0	−0.25	0.7491	0.1739	−0.5751	0.0587	0.0143	−0.0444
25	0.75	1	0.25	−0.7491	−1.3104	−0.5614	−0.0361	−0.0669	−0.0308
26	0.75	1	0.00	−0.0000	−0.5688	−0.5688	0.0000	−0.0316	−0.0316
27	0.75	1	−0.25	0.7491	0.1721	−0.5769	0.0417	0.0100	−0.0316

Appendix B

Observed Score-Based Differential Item Functioning Measures

Corollary 2.3 in Meredith and Millsap (1992) implies that under the null hypothesis ($b_r - b_f = 0$), the sufficient condition for $\Delta - \mathrm{DIF}_X = P - \mathrm{DIF}_X = 0$ to hold is X being the sufficient statistic for θ, no matter whether $\mu_r - \mu_f = 0$ or not (as shown with weighted sum scores). However, it may not be a necessary condition when $D = \mu_r - \mu_f = 0$ (as shown in the case of the 2PL model with the simple sum score).

For the 1PL model, Holland and Thayer (1988) showed that Equation 14 holds (i.e., $\Delta - \mathrm{DIF}_\theta = \Delta - \mathrm{DIF}_X = -4a(b_f - b_r)$) under both the null and alternative hypotheses, because the simple sum score X is the sufficient statistic for θ. Using their notations and arguments, we can easily show that Equation 14 is true for the 2PL model for the weighted sum score X^*.

More specifically, let $\mathbf{x} = (x_1, x_2, \cdots, x_J)$ be the item response vector, the likelihood function

$$p(\mathbf{x}) = \int \prod_{k=1}^{J} P_k(\theta)^{x_k} Q_k(\theta)^{1-x_k} \, dG(\theta),$$

where $G(\theta)$ is the cumulative distribution of θ, and

$$P_k(\theta) = P\left(Y_k = 1|\theta\right) = \frac{\exp\left(D_0 a_k\left(\theta - b_k\right)\right)}{1 + \exp\left(D_0 a_k\left(\theta - b_k\right)\right)},$$

$D_0 = 1.7$, and $Q_k(\theta) = 1 - P_k(\theta)$. Theorem 1 in Cressie & Holland (1983) shows that $p(x)$ can be expressed as

$$p(\mathbf{x}) = p(0) \int \prod_k \left[\frac{P_k(\theta)}{Q_k(\theta)}\right]^{x_k} dG(\theta) = p(0) \prod_k f_k^{x_k} \mu\left(D_0 x^*\right)$$

for some $p(0) > 0$, where $f_k = \exp\left(-D_0 a_k b_k\right)$, $x^* = \sum_{k=1}^{J} a_k x_k$, and $\mu(t) = \int \exp(t\theta) dG(\theta)$.

Assume that the studied item is the first item on the test and that the rest of the items are the same for the focal and reference groups, and assume $P\left(X^* = x^*\right) \neq 0$. Then, for the reference group R,

$$\begin{aligned} p_{R1} &= P\left(X_1 = 1|X^* = x^*, R\right) \\ &= \frac{P\left(X_1 = 1, X^* = x^*, R\right)}{P\left(X^* = x^*, R\right)} = \frac{f_{1R} S_{J-1,x^*-a_1}\left(f_R^*\right)}{S_{J,x^*}\left(f_R\right)}, \end{aligned}$$

where $f_R = \left(f_{1R}, f_R^*\right)$, $f_R^* = \left(f_{2R}, \cdots, f_{JR}\right)$, and

$$S_{J,x^*}(f) = \sum_{x: \sum_k a_k x_k = x^*} \prod_{k=1}^{J} f_k^{x_k}.$$

Similarly,

$$q_{R1} = 1 - p_{R1} = \frac{S_{J-1,x^*}\left(f_R^*\right)}{S_{J,x^*}\left(f_R\right)}.$$

Hence

$$\frac{p_{R1}}{q_{R1}} = f_{1R} \frac{S_{J-1,x^*-a_1}\left(f_R^*\right)}{S_{J-1,x^*}\left(f_R^*\right)}.$$

We can obtain, in the same fashion for the focal group F, that

$$\frac{p_{F1}}{q_{F1}} = f_{1F} \frac{S_{J-1,x^*-a_1}\left(f_F^*\right)}{S_{J-1,x^*}\left(f_F^*\right)}.$$

Then, the log odds ratio is

$$\frac{p_{R1} q_{F1}}{q_{R1} p_{F1}} = \frac{f_{1R}}{f_{1F}} = \exp\left(D_0 a_1\left(b_f - b_r\right)\right),$$

so $\Delta - \mathrm{DIF}_X = -4a_1(b_f - b_r)$.

Suggested citation:

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

Action Editor: Shelby Haberman

Reviewers: Rebecca Zwick and John Mazzeo