Research Report

A Note on Using Weighted Sum Scores in the P-DIF Statistic

ETS RR–19-32

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

A Note on Using Weighted Sum Scores in the P-DIF Statistic

Hongwen Guo & Neil J. Dorans

Educational Testing Service, Princeton, NJ

The Mantel–Haenszel delta difference (MH D-DIF) and the standardized proportion difference (STD P-DIF) are two observed-score methods that have been used to assess differential item functioning (DIF) at Educational Testing Service since the early 1990s. Latent-variable approaches to assessing measurement invariance at the item level have been proposed and studied since then as well. Previous research showed that using the weighted sum score as the matching variable may close the gap between the MH D-DIF statistic based on observed scores and its counterpart based on latent ability. In this study, we show that weighting reduces the difference between the STD P-DIF statistic and its counterpart based on latent ability. In addition, we discuss the factors that influence the gap and examine approaches that may facilitate the use of weighted STD P-DIF.

Keywords DIF; measurement invariance; matching variable; IRT; computation

doi:10.1002/ets2.12268

Differential item functioning (DIF) analyses are routinely performed in many testing programs. For dichotomous test items, the Mantel–Haenszel delta difference (MH D-DIF) statistic (Holland & Thayer, 1988) and the standardized proportion difference (STD P-DIF) statistic (Dorans & Kulick, 1986) are widely used at Educational Testing Service to screen items for DIF. Both methods use the observed score as the matching variable, and the observed score is usually the simple sum score. Differences have been observed in many simulation studies between the DIF statistics based on observed scores and their counterparts based on latent ability (Donoghue, Holland, & Thayer, 1993; Roussos, Schnipke, & Pashley, 1999; Zwick, 1990).

In recent studies (Guo & Dorans, 2019a, 2019b), we showed theoretically and analytically that using the weighted sum score as the matching variable may close the gap between the MH D-DIF statistic and its counterpart based on latent ability. In this paper, we show that using the weighted sum score as the matching variable can reduce the gap between the STD P-DIF statistic and its counterpart based on latent ability.

Two perspectives on assessing fairness at the item level have coexisted since the early 1990s: One reflects a belief in the practical advantages and conceptual clarity of the observed-score approaches; another reflects a belief that a model-based approach is more principled, flexible, and powerful (Holland & Wainer, 1993). We referred to the DIF based on true scores or latent ability as a departure from measurement invariance (DMI) to clarify that DIF and DMI measure departures from different definitions of invariance (Guo & Dorans, 2019b). The mathematical definitions of DMI and DIF are presented in the next section. In this study, we investigate analytically differences between STD P-DIF and its DMI counterpart. We focus on the uniform DMI case when item difficulty parameters are different for the focal and reference groups, but item discrimination parameters are the same.

More specifically, we analytically demonstrate that DIF and DMI measure different quantities under the uniform DMI assumption for the two-parameter logistic (2PL) item response theory (IRT) model and that use of DIF measures based on observed scores with a weighted sum score, instead of an unweighted sum, agree with the DMI measure. We also investigate three binning methods (naive, linear, and equipercentile) that could facilitate potential operational use of STD P-DIF with weighted sum scores.

Research Questions

In this section, we formally introduce the research questions and necessary notation and definitions. Let Y stand for the dichotomous item score on an item ($Y = 1$ when the item response is correct, and $Y = 0$ when the item response

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

is incorrect), X be the sum score on the test, and θ be the latent ability of the test taker that the test measures. The 2PL IRT model assumes that the probability of obtaining a correct response on item j for a test taker with the latent ability θ is

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

where $D_0 = 1.7$, and a_j and b_j are the item discrimination and difficulty parameters (Lord, 1980). When $a_j \equiv a$, the 2PL model reduces to a one-parameter logistic (1PL) model.

To assess whether an item functioned differently for test takers in two different subgroups, a focal group (Group f) and a reference group (Group r), a comparison can be made between $P_f(Y = 1|X)$ and $P_r(Y = 1|X)$ or between $P_f(Y = 1|\theta)$ and $P_r(Y = 1|\theta)$. The STD P-DIF statistic is a weighted difference of conditional P^+ values between the focal and reference groups for an item:

$$\text{STD P-DIF} = \sum_x w_{xf} P^+_{xf} - \sum_x w_{xf} P^+_{xr}, \tag{2}$$

where, for the stratum/matching variable (typically the total test score X), w_{fx} is the proportion of the focal group members in stratum $X = x$, and P^+_{xf} and P^+_{xr} are the average item scores for the focal and reference groups in stratum $X = x$, respectively. Note that for a number-right score, the index x can range from 0 to the number of items. The STD P-DIF statistic is an estimate of the expected difference between the item response functions, $P_f(X)$ and $P_r(X)$, with respect to the focal group; that is,

$$P\text{-DIF}_X = \sum_x \left[P_f(x) - P_r(x)\right] g_f(x), \tag{3}$$

where $P(x) = P(Y = 1|x)$ and $g_f(x)$ is the probability distribution of x for the focal group.

Alternatively, if the latent variable were known, a criterion based on latent ability $P\text{-DMI}_\theta$ might be calculated by matching the item scores on the latent variable θ:

$$P\text{-DMI}_\theta = \int_\theta \left\{P_f(\theta) - P_r(\theta)\right\} \psi_f(\theta)\, d\theta, \tag{4}$$

where $P(\theta) = P(Y = 1|\theta)$ and $\psi_f(\theta)$ is the density function of θ for the focal group. The P-DMI measures the expected difference between the item response functions, $P_f(\theta)$ and $P_r(\theta)$ with respect to the focal group. Equation 4 is the quantity that SIBTEST (Chang, Mazzeo, & Roussos, 1996; Shealy & Stout, 1993) attempted to estimate via subgroup-specific linear transformations of observed scores (Dorans, 2011).

The research questions we address in this study are

1. When do $P\text{-DMI}_\theta$ and $P\text{-DIF}_X$ return the same value under the 1PL and 2PL models?
2. What factors play important roles in producing differences between $P\text{-DMI}_\theta$ and $P\text{-DIF}_X$?
3. Can a modification of $P\text{-DIF}_X$ reduce the difference between it and its DMI counterpart?
4. How well do calculations based on binned total scores approximate the modified $P\text{-DIF}_X$?

As in Guo and Dorans (2019a, 2019b), we use population distributions and IRT models to build our arguments; no item responses are involved (that is, we do not use any simulated or real data).

Assumptions and Methods

To address these research questions, we analytically computed the quantities (P-DIF and P-DMI) under the following model assumptions:

- The latent ability is unidimensional, item responses are independent given the latent ability, and the item response function is of the form in Equation 1.
- The studied item may exhibit uniform DMI, but the rest of the items exhibit no DMI.

Computation

The computational procedures for $P\text{-DIF}_X$ for either simple sum or weighted sum scores use a combination of different algorithms, including the recursion formula in Lord and Wingersky (1984) to obtain conditional probabilities given the latent variable, Bayesian formulas to obtain conditional probabilities given simple sum scores, and convolution of independent random variables to obtain conditional probabilities with weighted sum scores. In particular,

- The recursion formula in Lord and Wingersky (1984) produces the conditional score distribution $f(x|\theta) = f(X = x|\theta)$ for given a θ value, where X is the simple sum score variable.
- Integration of $f(x|\theta)$ over θ produces the unconditional distribution $g(x)$ of X, needed in Equation 3; that is, $g(x) = \int_\theta f(x|\,\theta)\psi(\theta)d\theta$, where $\psi(\theta)$ is the ability density distribution.
- The conditional distribution, $P(Y = 0|\,X = x) = \int_\theta \{[P(X = x|\,Y_j = 0, \theta) \times P(Y = 0|\,\theta)]/f(x|\,\theta)\psi(\theta)\}d\theta$, is needed in Equation 3 as well.
- For given θ, to find the conditional distribution of the weighted sum score $Z = w_1Z_1 + w_2Z_2$, where w_i, $i = 1, 2$ are constants, and Z_1 and Z_2 are independent variables, the convolution technique yields

$$P(Z = z|\theta) = \int_t P\left(w_1Z_1 = t|\theta\right) \times P\left(w_2Z_2 = z - t|\theta\right) dt.$$

For more detailed information, interested readers can refer to Guo and Dorans (2019a, pp. 4–6).

Binning

In this subsection, we introduce three binning methods that could be used to facilitate the potential operational use of our theoretical results.

Let $X^* = \sum_{j=1}^{J} a_j y_j$ be the weighted sum of item scores, including the studied item, where a_j is the item discrimination parameter for item j, and $\mathbf{y} = (y_1, y_2, \cdots, y_J)$ is the item response vector on a test of J items. The weighted sum score can assume a very large number of values when the discrimination parameters differ across items, and matching on the exact weighted sum is challenging. Thus, binning is necessary for practical use of DIF when the weighted sum score is the matching variable.

The three binning methods are termed the *naive, linear*, and *equipercentile* methods (Guo & Dorans, 2019b). In the *naive* matching method, the weighted score range 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 one interval (bin) are matched to those in the same score interval in the reference group (note that this is equivalent to rounding or truncating the weighted sum score to an integer value when the score ranges are the same as the unweighted sum score). For the other two methods, we used the linear and equipercentile (EQ) equating methods of test scores to find the boundaries between bins (Guo & Dorans, 2019b; Kolen & Brennan, 2004). More specifically, let X and Z be the possible simple sum score and weighted sum score. For the *linear* binning method, for each possible sum score $i \in \{0, 1., \ldots, J\}$, we find $z_i = \frac{\sigma_z}{\sigma_x}\left(i - \mu_x\right) + \mu_z$ and then create bins for the weighted sum scores defined by the intervals $(z_i - H, z_i + H]$, where μ_x, μ_z, σ_x, and σ_z are the score means and standard deviations, and $H = \sigma_z/(2\sigma_x)$. Similarly for the *equipercentile* binning method, we define $z_i = F_z^{-1}\left[F_x(i)\right]$, where $F_X(x)$ and $F_Z(z)$ are the corresponding cumulative score distributions, respectively. Test takers whose weighted scores are in the same interval of $(z_x - H_x, z_x + H_x)$ are considered to be matched, where $H_x = (z_x - z_{x-1})/2$. The probability of Z in one bin is replaced by the sum of individual probabilities of Z in that bin; that is

$$P(z \in \left(z_x - H_x, z_x + H_x\right]) = \sum_{z \in (z_x - H_x, z_x + H_x]} P_z.$$

Test Design

Under the 1PL model, the difference between DIF and DMI will diminish as the test length increases (Holland & Thayer, 1988). We focus on the 2PL model with an ability difference between the focal and reference groups, because this is where a large discrepancy is likely to be observed between the STD P-DIF statistics and the corresponding DMIs (Guo & Dorans,

Table 1 Item Discrimination and Difficulty Parameters Serving as Core Items for Our Illustrations

Item	Discrimination (a)	Difficulty (b)
1	.48	−1
2	.48	0
3	.48	1
4	.60	−1
5	.60	0
6	.60	1
7	.75	−1
8	.75	0
9	.75	1

2019a; Zwick, 1990). It was observed that when the two groups do not differ in ability, differences between DIF and DMI are negligible (Guo & Dorans, 2019a). Based on previous studies (Camilli & Shepard, 1994; Chang et al., 1996; Zwick, Thayer, & Mazzeo, 1997) and test information from a large-scale testing program, we designed the study as follows. For easy interpretation of the results, we decided to use three levels of item difficulty and three levels of item discrimination that, when crossed, would give us a core of nine items that span a realistic range. These levels are:

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

Table 1 contains the item parameter combinations of the nine-item core that were used to create our hypothetical *tests*. We constructed a 27-item hypothetical test by using this nine-item core three times. We also constructed a 54-item test by using this nine-item core six times.

In our test design, we varied test length to study its effect on the agreement between DIF and DMI measures:

- Test length: The number of items on the test was either $J = 27$ (a short test) or $J = 54$ (a medium-length test).

In addition to crossing difficulty and discrimination, we varied the degree of departure from measurement invariance:

- d size: The difference in item difficulties for the studied item between the focal and reference groups $d = b_f - b_r$ was set to be −.25, 0, or .25 (the item discrimination parameters were the same for the two groups so that the design represented a uniform DIF case).

For the test length of $J = 27$, there were 27 versions of the test, defined by crossing a (3 levels), b_r (3 levels), and d (3 levels) for the studied item. Each version of the test differed by the a, b_r, and d values of the studied item.

For example, Table 2 shows the first three versions of the test. Version 1 of the test focuses on Item 1, which has $a = .48$, $b_r = -1$, and $d = .25$; the rest of the items from Item 2 to Item 27 have item parameters listed below Item 1 and are free of DMI. Version 2 focuses on Item 2, which has $a = .48$, $b_r = -1$, and $d = 0$; the rest of the items (Item 1, Item 3 to Item 27) have item parameters listed in the sixth and seventh columns and are free of DMI. Version 3 focuses on Item 3, which has $a = .48$, $b_r = -1$, and $d = -.25$; the rest of the items (Item 1, Item 2, and Item 4 to Item 27) have item parameters listed in the 10th and 11th columns and are free of DMI.

Table 3 lists all the 27 versions in a condensed format. Each row shows a test version, the corresponding item parameters are those of the studied item, and the remaining 26 items on this version have item parameters listed in the second and third columns and are free of DMI.

On nine of the 27 versions (i.e., Versions 2, 5, 8, 11, 14, 17, 20, 23, 26) of the test, the studied item was free of DMI. These nine versions were psychometrically parallel to each other with respect to item difficulty, item discrimination, and DMI; they differed only in the studied item (i.e., these nine versions are the null conditions). Of the remaining 18 versions, nine (i.e., Versions 1, 4, 7, 10, 13, 16, 19, 22, 25) had a studied item with positive DMI, and nine (i.e., Versions 3, 6, 9, 12, 15, 18, 21, 24, 27) had a studied item with negative DMI.

Table 2 The First Three Versions of the P-DIF Tests

Version 1				Version 2				Version 3			
Item	a	b_r	d	Item	a	b_r	d	Item	a	b_r	d
1	**.48**	**−1**	**.25**	1	.48	−1		1	.48	−1	
2	.48	−1		**2**	**.48**	**−1**	**0**	2	.48	−1	
3	.48	−1		3	.48	−1		**3**	**.48**	**−1**	**−.25**
4	.48	0		4	.48	0		4	.48	0	
5	.48	0		5	.48	0		5	.48	0	
6	.48	0		6	.48	0		6	.48	0	
7	.48	1		7	.48	1		7	.48	1	
8	.48	1		8	.48	1		8	.48	1	
9	.48	1		9	.48	1		9	.48	1	
10	.60	−1		10	.60	−1		10	.60	−1	
11	.60	−1		11	.60	−1		11	.60	−1	
12	.60	−1		12	.60	−1		12	.60	−1	
13	.60	0		13	.60	0		13	.60	0	
14	.60	0		14	.60	0		14	.60	0	
15	.60	0		15	.60	0		15	.60	0	
16	.60	1		16	.60	1		16	.60	1	
17	.60	1		17	.60	1		17	.60	1	
18	.60	1		18	.60	1		18	.60	1	
19	.75	−1		19	.75	−1		19	.75	−1	
20	.75	−1		20	.75	−1		20	.75	−1	
21	.75	−1		21	.75	−1		21	.75	−1	
22	.75	0		22	.75	0		22	.75	0	
23	.75	0		23	.75	0		23	.75	0	
24	.75	0		24	.75	0		24	.75	0	
25	.75	1		25	.75	1		25	.75	1	
26	.75	1		26	.75	1		26	.75	1	
27	.75	1		27	.75	1		27	.75	1	

As noted above, the test length $L = 54$ was created from six parallel versions of the nine-item core depicted in Table 1. Similarly, there were 27 versions of the test, and only one item on each version served as the studied item. The remaining 53 items on each of the 27 versions were free of DMI. For each of the 27 versions, the location of the studied item and its DMI value varied. The properties for the studied item for each of the 27 versions of the 54-item test are also presented in Table 3.

Finally, the ability difference between the focal and reference groups was one standard deviation unit: The reference group ability followed a normal distribution with mean $\mu_r = .50$ and standard deviation $\sigma_r = 1$, and the focal group ability also followed a normal distribution with mean $\mu_f = -.50$ and standard deviation $\sigma_f = 1$, to represent cases when there were large group differences (i.e., $\mu_f - \mu_r = 1$).

Note that the large ability difference and the simplified test design were chosen to facilitate analytic computation.

Analytic Results

As noted above, in each of the analyses below, only the studied item differs in the b parameter between the focal and reference groups; the remaining items stay the same for the two groups. For instance, in Table 4, the first row shows the DIF measures on one test version when the first item ($a_1 = .48$, $b_{1,r} = -1$, $b_{1,f} = b_{1,r} + d_1 = -1 + .25 = -.75$) differs in item difficulty for the focal and reference groups, while the remaining 26 items are the same across the two groups. The second row presents the results for the second item where $d = 0$ (i.e., $a_1 = .48$, $b_{1,r} = -1$, and $b_{1,f} = b_{1,r}$). Note again, the DIF measures are not computed from any item response data; instead, they are derived from the assumptions in the Assumptions and Methods section and the test design in the Test Design section.

In the following, we present the results for the short test ($J = 27$) first, and then we present those for the longer test ($J = 54$).

Table 3 The 27 Versions of the Test Defined by Crossing the a, b_r, and d Values for the Studied Item

Version of test	Studied item discrimination (a)	Studied item difficulty (b)	Studied item DMI (d)
1	.48	−1	.25
2	.48	−1	0
3	.48	−1	−.25
4	.48	0	.25
5	.48	0	0
6	.48	0	−.25
7	.48	1	.25
8	.48	1	0
9	.48	1	−.25
10	.60	−1	.25
11	.60	−1	0
12	.60	−1	−.25
13	.60	0	.25
14	.60	0	0
15	.60	0	−.25
16	.60	1	.25
17	.60	1	0
18	.60	1	−.25
19	.75	−1	.25
20	.75	−1	0
21	.75	−1	−.25
22	.75	0	.25
23	.75	0	0
24	.75	0	−.25
25	.75	1	.25
26	.75	1	0
27	.75	1	−.25

Note. DMI = departure from measurement invariance.

The Short Test ($J = 27$)

Table 4 shows the P-DIF measures obtained from different computation methods: PDMI_θ is the P-DMI size based on θ; PDIF.SS is the P-DIF measure based on observed scores, matching by the simple sum score X; PDIF.WS is the P-DIF measure based on observed scores, matching by the weighted sum score with exact matching; and PDIF.WS.naive, PDIF.WS.linear, and PDIF.WS.eq are the DIF measures based on observed scores, matching by the weighted sum score with naive, linear, and equipercentile binning methods.

As shown in Table 4, PDIF.SS, matched on the simple sum, produced the largest difference under both the null hypothesis $d = 0$ and the alternative hypothesis $d \neq 0$, except when $a = .60$. The difference was positive for $a = .48$ and negative for $a = .75$.

On the other hand, PDIF.WS, the exact matching with the weighted sum score, returned a zero under the null hypothesis of $d = 0$. Under the alternative hypothesis, the difference between PDMI_θ and PDIF.WS was greatly reduced when $a \neq .60$, compared to those for PDIF.SS, and the magnitude of this difference increased as a increased. In addition, as the relative item difficulty $|b_r + d - \mu_f|$ increased, $|P\text{-DIF}_X - P\text{-DMI}_\theta|$ decreased.

None of the binning methods in Table 4 returned a zero under the null hypothesis. However, the difference was less than 15% of that produced by PDIF.SS no matter whether DMI existed or not for $a \neq .60$. For $a = .60$, this difference was smaller as well compared to that produced by PDIF.SS. Note that under the alternative hypothesis, all the binning methods returned PDIF measures that were closer to $P\text{-DMI}_\theta$ than the exact matching method. The three binning methods were very similar; on average, the difference was larger when a was larger for all three matching methods.

Figure 1 portrays differences between P-DMI_θ and each of the P-DIF measures based on weighted sum scores: exact matching (△), naive binning (⋄), linear binning (+), and equipercentile binning (×). The figure is partitioned into nine panels. Each panel shows difference results for three test versions for one of the nine items in the core set; only d varies across each panel. The difference between P-DMI and P-DIF.SS is labeled as a circle (∘). These circles (∘) are absent from the first and third rows of the panels because their large differences are out of range for $a \neq .6$. The layout of Figure 1

Table 4 P-DMI$_\theta$ and Various P-DIF Measures for the Studied Item

Test version	a	b_r	d	PDMI$_\theta$	PDIF. SS	PDIF. WS	PDIF. WS. naive	PDIF. WS. linear	PDIF. WS. eq
1	.48	−1	.25	−.044	.004	−.039	−.044	−.042	−.042
2	.48	−1	0	0	.043	0	−.002	0	.001
3	.48	−1	−.25	.043	.081	.038	.040	.042	.042
4	.48	0	.25	−.043	.003	−.039	−.043	−.041	−.041
5	.48	0	0	0	.042	0	−.001	0	0
6	.48	0	−.25	.044	.081	.040	.041	.043	.043
7	.48	1	.25	−.033	.003	−.031	−.033	−.032	−.032
8	.48	1	0	0	.034	0	−.001	0	0
9	.48	1	−.25	.036	.067	.033	.034	.035	.035
10	.6	−1	.25	−.051	−.047	−.046	−.050	−.050	−.049
11	.6	−1	0	0	.002	0	−.001	−.001	0
12	.6	−1	−.25	.049	.049	.044	.046	.047	.047
13	.6	0	.25	−.049	−.045	−.045	−.049	−.048	−.048
14	.6	0	0	0	.002	0	−.001	0	0
15	.6	0	−.25	.051	.051	.046	.048	.049	.049
16	.6	1	.25	−.035	−.032	−.032	−.035	−.034	−.034
17	.6	1	0	0	.002	0	−.001	0	0
18	.6	1	−.25	.039	.039	.036	.037	.038	.038
19	.75	−1	.25	−.059	−.105	−.051	−.057	−.056	−.057
20	.75	−1	0	0	−.045	0	−.002	−.001	−.001
21	.75	−1	−.25	.056	.011	.048	.050	.051	.051
22	.75	0	.25	−.056	−.104	−.050	−.054	−.054	−.054
23	.75	0	0	0	−.048	0	−.002	−.001	−.001
24	.75	0	−.25	.059	.012	.053	.054	.054	.054
25	.75	1	.25	−.036	−.066	−.033	−.035	−.035	−.035
26	.75	1	0	0	−.032	0	−.001	−.001	−.001
27	.75	1	−.25	.042	.009	.038	.039	.039	.039

Note. Each row represents a different version of a 27-item test, in which the studied item is different for the focal and reference groups, but the other 26 items are always free of DMI. PDMI$_\theta$ is the P-DMI measure based on latent ability; PDIF.SS, PDIF.WS, PDIF.WS.naive, PDIF.WS.linear, and PDIF.WS.equ in columns 6 to 10 are P-DIF measures based on observed scores, matching by simple sum scores (SS), by weighted sum scores (WS) with exact matching, by WS with naive binning, by WS with linear binning, by WS with equipercentile binning, respectively. In each row, the studied item differs in b parameters (i.e., $b_f - b_r = d$) between the focal and reference groups, but the other items are the same.

assists comparison across three factors. In each panel of the figure, the DIF size factor was studied; in each column, the a-parameter effect was studied; and in each row, the b-parameter was studied.

More specifically, the first panel in Figure 1 has a studied item with $a = .48$ and $b = -1$, and the difference parameter $d = b_f - b_r$ is .25, 0, and − .25, corresponding to three different test versions. For exact matching (△), the differences from DMI are small in absolute value (<.005) and are positive, zero, and negative when $d = b_f - b_r$ is .25, 0, and − .25, respectively; those from the linear (×) and equipercentile (⋄) binning methods are smaller in absolute value as well (<.002); those (+) from the naive binning method are between .001 and .003 and are all negative.

Across each row in Figure 1, we observed that, as the item difficulty parameter b increases, the differences between P-DIF measures and P-DMI decrease. Across each column, we observed that the magnitude in the differences increases when a increases, except for d.ss when the matching variable in P-DIF is the simple sum score.

Overall, the item difficulty parameter b, the item discrimination parameter a, and the difference $d = b_F - b_R$ in the difficulty parameters all affect the differences between P-DIF and P-DMI; the two binning methods, linear and equipercentile, are comparable, and they are better than the naive method when items are not very easy. In addition, all the P-DIF measures based on weighted sum scores are comparable to or better than the best case (i.e., $a = .6$, the average item discrimination of the test) of P-DIF measure based on simple sum scores, when compared to the P-DMI criterion.

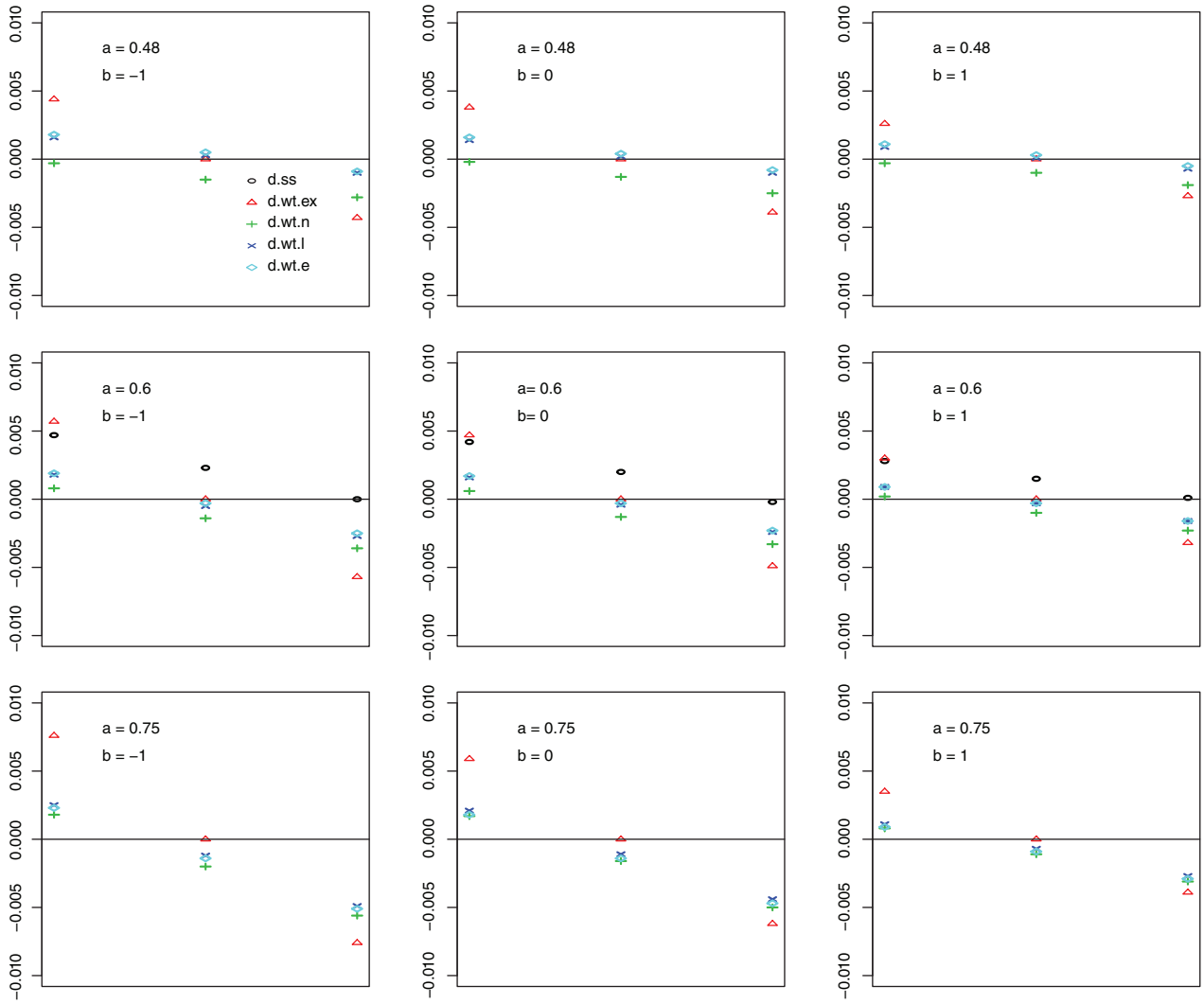

Figure 1 Differences between P-DMI and P-DIF measures: The difference between P-DIF.SS and P-DMI (d.ss., ∘) and differences between the P-DIF measures based on weighted scores and P-DMI (d.wt.ex. with exact matching, △; d.wt.n. with naive binning, ◇; d.wt.l. with linear binning, +; and d.wt.e. with equipercentile binning, ×) for each test version ($J = 27$). Note that the differences between P-DMI and P-DIF.SS are not shown in the first and third rows because of their large values.

The Longer Test ($J = 54$)

Table 5 shows the differences between P-DMI$_\theta$ and each of the different P-DIF measures for the test of $J = 54$, where the *a* parameters assume three values: .48, .60, and .75.

Again, the P-DIF measures with exact matching on weighted sum scores did not differ from P-DMI under the null hypothesis. In addition, the impact of the design factors (such as *a*, *b*, and *d*) on the difference between P-DMI and P-DIF measures for the longer test is similar to the impact shown in Table 4. All differences between P-DMI and P-DIF measures were reduced in the longer test, but the magnitude of the reduction was different for the P-DIF measures based on simple sum scores and weighted sum scores. Compared to Table 4, Table 5 shows that there is almost no reduction in the difference between DMI and DIF from $J = 27$ to $J = 54$ for the P-DIF.SS based on simple sum scores, but for the P-DIF based on weighted sum scores, the differences reduced by about 50% from the shorter test to the longer one. The difference produced by the P-DIF measures based on weighted scores is all in the third decimal places.[1]

Figure 2 displays the differences between P-DMI$_\theta$ and each of the P-DIF measures on 27 versions of the longer test ($J = 54$). We observed that when the test is longer, the differences among the three binning methods are smaller and hard

Table 5 P-DMI$_\theta$ and Various P-DIF Measures for the Studied Item

Test version	a	b_r	d	PDMI$_\theta$	PDIF SS	PDIF WS	PDIF WS & naive	PDIF WS & linear	PDIF WS & eq
1	.48	−1	.25	−.044	.003	−.043	−.044	−.043	−.043
2	.48	−1	0	0	.043	0	0	0	0
3	.48	−1	−.25	.043	.082	.041	.042	.042	.042
4	.48	0	.25	−.043	.003	−.042	−.042	−.042	−.042
5	.48	0	0	0	.042	0	0	0	0
6	.48	0	−.25	.044	.082	.043	.043	.043	.043
7	.48	1	.25	−.033	.002	−.033	−.033	−.033	−.033
8	.48	1	0	0	.034	0	0	0	0
9	.48	1	−.25	.036	.068	.035	.035	.035	.036
10	.6	−1	.25	−.051	−.048	−.050	−.051	−.050	−.050
11	.6	−1	0	0	.002	0	0	0	0
12	.6	−1	−.25	.049	.050	.048	.048	.048	.048
13	.6	0	.25	−.049	−.046	−.048	−.049	−.049	−.048
14	.6	0	0	0	.002	0	0	0	0
15	.6	0	−.25	.051	.052	.050	.050	.050	.050
16	.6	1	.25	−.035	−.033	−.034	−.035	−.035	−.035
17	.6	1	0	0	.002	0	0	0	0
18	.6	1	−.25	.039	.040	.038	.038	.039	.039
19	.75	−1	.25	−.059	−.107	−.056	−.057	−.057	−.057
20	.75	−1	0	0	−.045	0	−.001	0	0
21	.75	−1	−.25	.056	.012	.053	.054	.054	.054
22	.75	0	.25	−.056	−.105	−.054	−.055	−.055	−.055
23	.75	0	0	0	−.048	0	0	0	0
24	.75	0	−.25	.059	.014	.057	.057	.057	.057
25	.75	1	.25	−.036	−.067	−.035	−.036	−.035	−.035
26	.75	1	0	0	−.032	0	0	0	0
27	.75	1	−.25	.042	.010	.040	.040	.041	.041

Note. Each row represents a different version of a 54-item test, in which the studied item is different for the focal and reference groups, but the other 53 items are always free of DMI. PDMI$_\theta$ is the P-DMI measure based on latent ability; PDIF.SS, PDIF.WS, PDIF.WS.naive, PDIF.WS.linear, and PDIF.WS.equ in columns 6 to 10 are P-DIF measures based on observed scores, matching by simple sum scores (SS), by weighted sum scores (WS) with exact matching, by WS with naive binning, by WS with linear binning, and by WS with equipercentile binning, respectively.

to differentiate. In contrast, the differences between P-DMI$_\theta$ and PDIF.ss based on simple sum scores are still too large to fit into the plot except for $a = .6$.

Discussion

When a 2PL model is the appropriate model for item response data, the P-DIF measures differ from the P-DMI measures, when the matching variable is the simple sum score and when the focal and reference groups differ in ability. Based on the analytic results we presented in this paper, the most influential factor on the difference between P-DMI$_\theta$ and P-DIF$_X$, under the uniform DIF scenarios and a 2PL model, appeared to be variation of the item discrimination parameters.[2] When the studied item had a discrimination parameter close to the average discrimination of the test, the two measures, P-DMI$_\theta$ and P-DIF$_X$, exhibited a small difference. In contrast, when the studied item discrimination parameter was very different from the average discrimination, the two measures exhibited a large difference, even under the null hypothesis of measurement invariance. That is, larger variation in a led to larger differences between the DIF measures based on simple sum scores and those based on weighted sum scores. Less and less variation in a means that the 2PL model is converging toward a 1PL model, in which case P-DIF$_X$ is expected to equal P-DMI$_\theta$. The relative item difficulty for the focal group played a smaller role: When the item was relatively harder, P-DIF$_X - P$-DMI$_\theta$ was slightly smaller. This result is different from what was found in MH D-DIF measures where item difficulty did not influence the differences (Guo & Dorans, 2019a, 2019b).

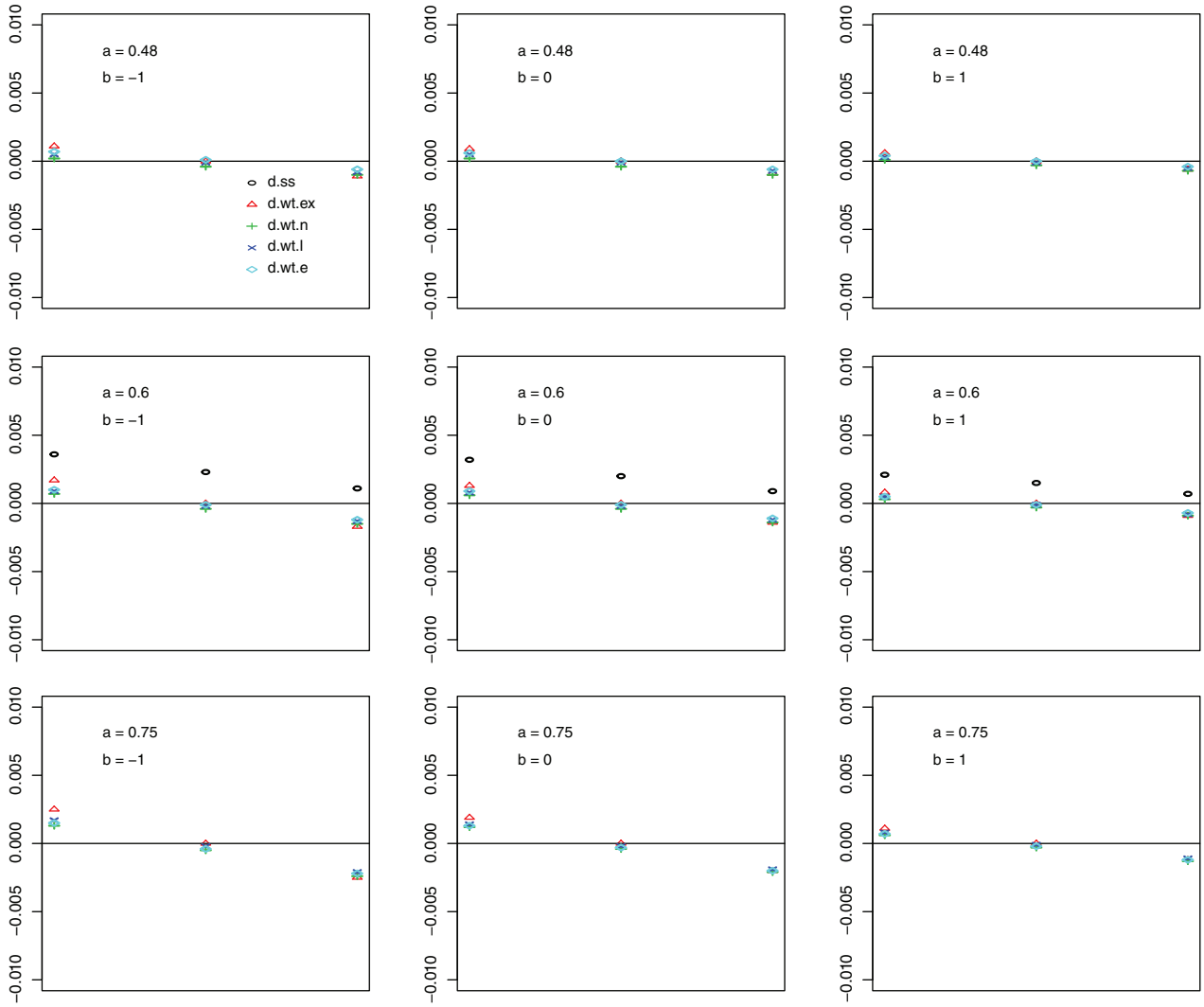

Figure 2 Differences between P-DMI and P-DIF measures: The difference between P-DIF.SS and P-DMI (d.ss., ∘) and between the P-DIF measures based on weighted scores and P-DMI (d.wt.ex. with exact matching, △; d.wt.n. with the naive binning, ◇; d.wt.l. with the linear binning, +; and d.wt.e. with the equipercentile binning, ×) for 27 versions of a longer test ($J = 54$). Note that the differences between P-DMI and P-DIF.SS are not shown in the first and third rows because of their large values.

In contrast, the difference between $P\text{-DMI}_\theta$ and $P\text{-DIF}_X$ with weighted sum scores was zero under the null hypothesis of no DMI. Even under the alternative hypothesis (when $d \neq 0$), the difference between $P\text{-DMI}_\theta$ and $P\text{-DIF}_X$ based on weighted sum scores was much smaller that observed between $P\text{-DMI}_\theta$ and $P\text{-DIF}_X$ based on simple sum scores.

In addition, binning methods may make it feasible to use the above methods in practice. While none of the binning methods returned a zero under the null hypothesis, they did greatly reduce the difference between $P\text{-DIF}_X$ based on simple sum scores and $P\text{-DMI}_\theta$. Although the linear and equipercentile binning methods performed slightly better than the naive method, the three binning methods were close to each other. We also noticed that $P\text{-DIF}_X$ based on weighted sum scores converged faster to $P\text{-DMI}_\theta$ measures than $P\text{-DIF}_X$ based on simple sum scores as the test length increased. We attributed this to the fact that the weighted sum scores became increasingly more reliable than simple sum scores. These observations were similar to those for the MH D-DIF measures (Guo & Dorans, 2019b).

Our results are derived analytically from model assumptions, rather than simulated item response data or real empirical data. More studies are necessary to investigate the impact of estimated item discrimination parameters, small sample sizes, and model fit to real data before we would recommend use of weighted P-DIF in practice. In addition, exploring how to use classical item statistics, such as biserial correlation (Lord, 1980), to weight item scores as an alternative to using IRT

parameter estimates might prove useful. Researchers may choose to use the estimated ability as the matching variable, but they need to address issues related to the sample size, model fit, and estimation error as well.

Acknowledgments

The authors would like to thank Rebecca Zwick and Tim Davey for their helpful comments and suggestions, and Kim Fryer and Ariel Katz for their editorial help. Any opinions expressed here are those of the authors and not necessarily those of Educational Testing Service.

Notes

1 The MH-type D-DIF measure seemed to converge to DMI faster than the P-DIF measures as the test became longer (Guo & Dorans, 2019b).

2 Actually, the most critical factor was the ability difference between the focal and reference groups, which led to $P\text{-DIF}_X = 0$ under the null hypothesis of measurement invariance. When the two groups had the same ability, the difference between $P\text{-DMI}_\theta$ and $P\text{-DIF}_X$ was negligible (Guo & Dorans, 2019a).

References

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

Chang, H. H., Mazzeo, J., & Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *Journal of Educational Measurement, 33*, 333–353. https://doi.org/10.1111/j.1745-3984.1996.tb00496.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). 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

Dorans, N. J. (2011). Holland's advice during the fourth generation of test theory: Blood tests can be contests. In N. J. Dorans & S. Sinharay (Eds.), *Looking back: Proceedings of a conference in honor of Paul W. Holland* (pp. 259–272). New York, NY: Springer-Verlag. https://doi.org/10.1007/978-1-4419-9389-2_14

Guo, H., & Dorans, N. (2019a). *Observed scores as matching variables in differential item functioning under the 1PL and 2PL model: Population results* (ETS Research Report RR-19-06). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12243

Guo, H., & Dorans, N. (2019b). *Using weighted sum scores to close the gap between DIF practice and theory*. Manuscript submitted for publication.

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. (2004). *Test equating, linking, and scaling: Methods and practices* (2nd ed.). New York, NY: Springer-Verlag. https://doi.org/10.1007/978-1-4757-4310-4

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

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

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

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

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., 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

Suggested citation:

Guo, H., & Dorans, N. J. (2019). *A note on using weighted sum scores in the P-DIF statistic* (Research Report No. RR-19-32). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12268

Action Editor: Rebecca Zwick

Reviewers: Tim Davey and Paul Jewsbury