Article

Applied Psychological Measurement
2018, Vol. 42(3) 206–220

DOI: 10.1177/0146621617726786
journals.sagepub.com/home/apm
SAGE

Improving the Assessment of Differential Item Functioning in Large-Scale Programs With Dual-Scale Purification of Rasch Models: The PISA Example

Cheng-Te Chen[1] and Bo-Sien Hwu[2]

Abstract

By design, large-scale educational testing programs often have a large proportion of missing data. Since the effect of missing data on differential item functioning (DIF) assessment has been investigated in recent years and it has been found that Type I error rates tend to be inflated, it is of great importance to adapt existing DIF assessment methods to the inflation. The DIF-free-then-DIF (DFTD) strategy, which originally involved one single-scale purification procedure to identify DIF-free items, has been extended to involve another scale purification procedure for the DIF assessment in this study, and this new method is called the dual-scale purification (DSP) procedure. The performance of the DSP procedure in assessing DIF in large-scale programs, such as Program for International Student Assessment (PISA), was compared with the DFTD strategy through a series of simulation studies. Results showed the superiority of the DSP procedure over the DFTD strategy when tests consisted of many DIF items and when data were missing by design as in large-scale programs. Moreover, an empirical study of the PISA 2009 Taiwan sample was provided to show the implications of the DSP procedure. The applications as well as further studies of DSP procedure are also discussed.

Keywords

differential item functioning, item response theory, scale purification, large-scale testing programs, missingness, balanced incomplete block design

Differential item functioning (DIF) assessment has become routine practice in real tests to ensure test fairness for examinees of different groups. Numerous DIF assessment methods have been proposed in the past several decades, including the Mantel–Haenszel (MH) method (Holland & Thayer, 1988), the logistic regression method (Swaminathan & Rogers, 1990), the

[1]National Tsing Hua University, Hsinchu, Taiwan
[2]National Sun Yat-sen University, Kaohsiung, Taiwan

Corresponding Author:
Cheng-Te Chen, Department of Educational Psychology and Counseling, National Tsing Hua University, No. 101, Section 2, Kuang-Fu Road, Hsinchu 30013, Taiwan.
Email: chengte@mx.nthu.edu.tw

simultaneous item bias test method (Shealy & Stout, 1993), the likelihood ratio test method (Thissen, Steinberg, & Wainer, 1988), and the multiple indicators, multiple causes confirmatory factor analysis method (Oort, 1998). The performance of these methods has been thoroughly investigated under various conditions through simulation studies, mostly with complete data. However, data are not always complete in reality. For example, there is always a considerable amount of missing data in large-scale programs such as the Program for International Student Assessment (PISA) and the National Assessment of Educational Progress program. This is because these large-scale programs use matrix sampling, in which every student answers only a small proportion of items, which leaves a large proportion of missingness by design in the response data. Because DIF assessment methods were not originally developed to assess DIF with missing data, the performance of these assessment methods in assessing DIF response data with a great deal of missing data by design has to be investigated.

The DIF assessment in large-scale programs has been conducted on variables such as gender, test language, immigration, as well as country (Grisay, Gonzalez, & Monseur, 2009; Organization for Economic Co-operation and Development [OECD], 2002, 2012). The PISA tests have used the equal-mean-difficulty (EMD) method for DIF assessment (OECD, 2002); however, this method has been shown to perform acceptably only when the test is perfect (i.e., no DIF items) or when the DIF magnitudes are balanced (Wang, 2004; Wang & Yeh, 2003). Because these two conditions may not always be the case in practice, the PISA needs other DIF assessment methods that are more suited to DIF variations.

Several researchers have conducted simulation studies to investigate the effect of missing data treatments and mechanisms on the performance of DIF analysis, for both uniform and nonuniform DIF (Finch, 2011; Robitzsch & Rupp, 2009). It is found that incorrectly treating data that are missing at random as incorrect can lead to sharp increases in Type I error rates and decreases in power rates in the MH method and the logistic regression method (Robitzsch & Rupp, 2009). Furthermore, several studies have been aimed at the effect of designed missingness on DIF assessment with the MH method (Allen & Donoghue, 1996; Goodman, Willse, Allen, & Klaric, 2011). To analyze structurally missing data due to balanced incomplete block (BIB) booklet designs, Sandilands (2014) included Lord's chi-square method (Lord, 1980) and compared its performance with the MH method. The results generally indicated that the Lord's chi-square method lost control of Type I error rates, whereas the MH method showed low power rates under the investigated conditions. Because the MH method and Lord's method use all-other-items (AOI) to form the matching variable, and both performed poorly, a more robust DIF assessment method that is able to construct an efficient matching variable would be the key to DIF assessment of data with a large proportion of designed missingness.

One recently developed DIF strategy, called the DIF-free-then-DIF (DFTD; Wang, 2008; Wang, Shih, & Sun, 2012), adopts a set of presumed constant-items (CI) as matching variables. Researchers have shown that the DFTD strategy provides well-controlled Type I error rates as well as satisfactory power on DIF assessment (J.-H. Chen, Chen, & Shih, 2014; Shih & Wang, 2009; Wang et al., 2012); hence, this strategy may also perform well on data with designed missingness. However, the major drawback of the DFTD strategy is that several DIF-free items first have to be located and anchored for the DIF assessment, so these anchor items are no longer tested for DIF in the later stage. As long as the anchors are not 100% DIF-free, which is the case in several studies (J.-H. Chen et al., 2014; Shih & Wang, 2009; Wang et al., 2012), a test of the DIF toward the anchor items should always be recruited. C.-T. Chen, Wang, and Shih (2012) piloted a modified DFTD strategy, called dual-scale purification (DSP), to overcome this problem by adding another scale purification procedure and applied the DSP to the assessment of differential ''rater'' functioning. This study focused on DIF assessment, which is much often conducted than the assessment of differential rater functioning and aimed to investigate

Table 1. List of Acronyms and Their Full Forms.

Acronym	Full form
AOI	All-other-item
AOI-P	All-other-item-with-purification
BIB	Balanced incomplete block
CI	Constant-item
DFTD	DIF-free-then-DIF
DIF	Differential item functioning
DSP	Dual-scale purification
EMD	Equal-mean-difficulty
ICI	Iterative-constant-item
IRT	Item response theory
MH	Mantel–Haenszel
OECD	Organization for Economic Co-operation and Development
PISA	Program for International Student Assessment

(a) whether the DFTD strategy would still perform well on data with designed missingness and, more importantly, (b) whether the DSP outperformed the DFTD on either complete or designed missingness data.

The rest of the article is organized as follows. First, the three methods (i.e., EMD, AOI, and CI) of building a common scale for DIF assessment, as well as the extension of the AOI method to a purification procedure, are introduced. Next, the DFTD strategy and the DSP procedure are explained. After the item response theory (IRT)-based DIF assessment methods are briefly introduced, the BIB booklet design used in the PISA is discussed, and an empirical example is given. Afterward, the series of simulations conducted in this study are described. Finally, conclusions and discussions are drawn, and some suggestions about DIF assessment on large-scale data are offered. A list of acronyms and the full forms are provided in Table 1 for reference.

Three Methods of Building a Common Scale for DIF Assessment

It is a prerequisite to place examinees from different groups onto a common scale in DIF assessment, regardless of whether the data are complete or are incomplete by design. There are several ways to build a common scale for groups of examinees: the EMD method (Wang, 2004), the AOI method, and the CI method (Wang & Yeh, 2003). In the EMD method, which has been adopted by the PISA for assessing DIF (OECD, 2002), the mean item difficulties are constrained to be identical across groups. However, by definition, such an assumption holds only when (a) the test does not contain any DIF items or (b) the test contains multiple DIF items and their DIF magnitudes are exactly balanced across groups. Because real tests are imperfect and mostly (if not always) consist of DIF items, these two conditions are rarely met in reality, making the EMD method unfeasible (Wang, 2004; Wang & Yeh, 2003). Other than the EMD method, the other way to build a common scale is to use a set of anchor items.

Two different anchor item methods have been proposed, which differ in the number of anchor items. In the AOI method, only the studied item (i.e., the one being assessed for DIF) is assumed to have DIF, while all other items in the entire test are assumed to be DIF-free (as anchor items). The IRT-based DIF detection with the likelihood ratio test has been taken as an example here. In a 10-item test, all 10 items are constrained to be DIF-free in the compact model; whereas in the augmented model, all nine items except the studied are constrained to be DIF-free and the studied item is allowed to have different parameters for different groups. In

the CI method, a constant set of items in a test are assumed to be DIF-free irrespective of studied items (Wang & Yeh, 2003). A practicer needs to select a set of DIF-free items as anchors. If he or she choose the first four items as anchors, in the compact model, only the first four items and the studied item are constrained to be DIF-free, whereas the other items are allowed to have different parameters for different groups. In the augmented model, only the first four items are constrained to be DIF-free, whereas all other items are allowed to have different parameters for different groups. The AOI method performs appropriately when the studied item is the only DIF item in the test, which is not always the case in real tests. If there are multiple DIF items in the test, the AOI method tends to yield inflated Type I error rates and therefore incorrect DIF assessment results. The more DIF items in the test, the more inflation in Type I error rates. However, the CI method can yield reasonable DIF assessment results given that the anchor items are indeed DIF-free. However, if the DIF items are misidentified as anchors, the DIF assessment results of the CI method might be incorrect (Wang & Yeh, 2003). In general, each of the three methods is flawed: the assumptions of the EMD method are rarely met in reality, the Type I error rates of the AOI method are inflated when there are many DIF items, a set of DIF-free items is required for the CI method.

To control the inflated Type I error rates in the AOI method when there are many DIF items, a scale purification procedure was proposed to cleanse the common scale of DIF items (Candell & Drasgow, 1988; Holland & Thayer, 1988). Within the IRT framework, the scale purification generally contains the following steps:

1. Assess Item 1 for DIF while treating all other items as anchors; assess Item 2 for DIF while treating all other items as anchors; repeat this step until the last item is assessed for DIF (i.e., the AOI method). Assume Items 1 and 2 in a 10-item test are deemed as DIF items in this step.
2. Remove those items deemed as exhibiting DIF in the previous step from the anchors and assess DIF for all items in the test again. For example, when detecting Item 1 or 2 for DIF, the updated anchors consist of Items 3 to 10. When detecting Item 3, the anchors consist of Items 4 to 10; when detecting Item 4, the anchors consist of Items 3, 5 to 10; and so on for Item 10.
3. Repeat Step 2 until the same set of items is identified as having DIF in two successive iterations (Steps 2 and 3 are the purification procedure).

Steps 1 to 3 are referred to as the all-other-item-with-purification (AOI-P) method. It has been found that the AOI-P method outperforms DIF assessment methods without scale purification in terms of Type I error rates and power rates when the studied test consists of multiple DIF items. However, when there are too many DIF items (e.g., more than 20% of items have DIF), the purification procedure fails from the very first step and yields inflated Type I error rates and deflated power rates (Wang et al., 2012). In general, the purification procedure only works when the common scale is not severely polluted.

The DFTD Strategy

Real tests may consist of more than 20% of DIF items. It is challenging to accurately assess DIF when tests contain many DIF items. Acknowledging that a common scale should comprise exclusively DIF-free items to yield accurate DIF assessment, Wang and his associates proposed the DFTD strategy for DIF assessment (Wang, 2008; Wang et al., 2012), which involves two steps. First, in the ''DIF-free-item-searching'' step, a set of items that is most likely to be DIF-free is selected as anchor for building a common scale. Second, in the ''DIF-testing'' step, all

but the anchor items are then assessed for DIF with the CI method, using those items selected in the first step as anchors. Identifying a set of DIF-free items to serve as anchors is critical in the DFTD strategy (Kopf, Zeileis, & Strobl, 2015; Woods, 2009). Instead of assessing DIF iteratively using scale purification procedures, the first step of the DFTD strategy is to locate a small set of DIF-free items to serve as anchors for the next step of DIF assessment.

As proposed by Wang (2008) and implemented by Shih and Wang (2009) and Wang et al. (2012), the iterative-constant-item (ICI) method could be used to identify DIF-free items in the DIF-free-item-searching step, which consists of the following three steps:

1. Set Item 1 as an anchor and assess the DIF of the other items; set Item 2 as an anchor and assess the DIF of the other items; continue until the last item is set as an anchor.
2. Sum the absolute values of the DIF amounts across the iterations for each item.
3. Select a prespecified number of items (e.g., four or eight items) that have the smallest sum to serve as anchors.

After the anchors are selected, the other items are tested for DIF in the DIF-testing step. Even though choosing more DIF-free items to serve as anchors generates commensurately higher powers in DIF assessments, this comes at a price: The risk that DIF items will be chosen mistakenly to serve as anchors also increases. Fortunately, it has been found that four DIF-free items are enough to yield a satisfactory power (Shih & Wang, 2009; Thissen et al., 1988; Wang, 2004; Wang, 2008; Wang & Yeh, 2003; Woods, 2009).

The DSP Procedure

Although simulation studies have shown that the DFTD strategy outperforms traditional scale purification methods such as the AOI-P (Wang et al., 2012), the ICI method in the DIF-free item-searching step is not able to locate exclusively DIF-free items. This warranted an assessment of anchor items for DIF in the DFTD strategy. C.-T. Chen et al. (2012) made such an attempt in the assessment of differential ''rater'' functioning. In their revision of the DFTD strategy, DSP procedures were adopted by replacing the CI method in the DFTD with the AOI-P method. In this study, the authors applied the DSP procedure and implemented it for the DIF assessment. To illustrate the DSP procedure, a 10-item test was assessed for DIF. Specifically, the DSP procedure consists of the following steps:

1. Use the ICI method to locate a set of DIF-free items (e.g., four items) to serve as anchors (the first-scale purification). Let us assume that Items 1 to 4 are identified as DIF-free in this step.
2. Assess all items for DIF with the AOI method using the anchors identified in the previous step. In this case, when Item 1 is assessed for DIF, Items 2 to 4 serve as the anchors; when Item 2 is assessed, Items 1, 3, and 4 serve as the anchors; when Item 3 is assessed, Items 1, 2, and 4 serve as the anchors; when Item 4 is assessed, Items 1 to 3 serve as the anchors; when Items 5 to 10 are assessed, Items 1 to 4 serve as the anchors.
3. Use those items deemed as DIF-free in the previous step as the anchors to assess all items for DIF again. In this case, let us assume that Items 3 to 8 are deemed as DIF-free in the previous step. When assessing Items 1, 2, 9, and 10 (nonanchored items) for DIF, Items 3 to 8 serve as the anchors. When Item 3 is assessed, Items 4 to 8 serve as the anchors; when Item 4 is assessed, Items 3, 5 to 8 serve as the anchors; and so on for assessing Items 5 to 8.
4. Repeat Step 3 until the same set of DIF items are located (the second-scale purification).

In the DSP procedure, an additional scale purification procedure is incorporated, which will generally have more items as anchors than the DFTD strategy and is thus anticipated to yield a higher power. Until now, the DFTD strategy has been investigated only with complete data. How it would perform in large-scale programs (having a large proportion of missingness) is unknown. Furthermore, the potential superiority of the DSP procedure over the DFTD strategy needs to be evaluated. The two questions will be answered with simulations.

IRT Models for DIF Assessment

To be in line with the use of models in the PISA (OECD, 2002), the authors concentrated on the Rasch model in this study; however, the general idea applies to other IRT models. For dichotomous items, one may fit the simple logistic Rasch model (Rasch, 1960):

$$\log\left(\frac{P_{ni1}}{P_{ni0}}\right) = \theta_n - \delta_i, \tag{1}$$

$$\theta_n \sim N\left(\mu, \sigma^2\right), \tag{2}$$

where P_{ni1} and P_{ni0} are the probability of person n answering item i correctly and incorrectly, respectively; θ_n is person n's latent trait and is often assumed to follow a normal distribution with a mean μ and variance σ^2; and δ_i is item i's level of difficulty. To identify the model, usually either the mean of the latent trait or the mean of the item difficulties is constrained at zero.

In the Rasch framework, an item is deemed to exhibit DIF when it has different parameters for different groups of persons. As Proposed by Paek and Wilson (2011), the Rasch DIF model is

$$\log\left(\frac{P_{nig1}}{P_{nig0}}\right) = \theta_n - \left(\delta_i + A_g + \tau_{ig}\right), \tag{3}$$

$$\text{with} \quad \theta_n \sim N\left(0, \sigma^2\right), \tag{4}$$

where g stands for group membership ($g = 1, \ldots, G$), and different means (i.e., A_g) for different groups are assumed with a common variance on the latent trait (i.e., σ^2). For identification, one of the groups has to be set as the reference group and related parameters would not be estimated. Therefore, δ_i is now item i's level of difficulty for the reference group. The parameter τ_{ig} is the group-specific difference from the reference, known as the DIF parameter. If the null hypothesis of $\tau_{ig} = 0$ is statistically rejected, then item i is deemed as exhibiting DIF.

The BIB Designs

In large-scale assessment, there are normally too many items for a test-taker to complete. As a result, items are allocated to booklets, and there are small proportions of common items between booklets so that all items are linked and placed on a common scale. According to the technical report of the PISA 2009, a total of 188 items were allocated to 13 separate item clusters (three clusters for the math subject, three clusters for the science subject, and seven clusters for the reading subject) and 13 booklets. Each booklet had four clusters. The BIB design was used to allocate clusters to booklets so that each cluster appeared in each of the four possible positions within a booklet once, and each pair of clusters appeared in one and only one booklet (OECD, 2012). Table 2 illustrates this practice with 13 clusters in 13 booklets. These booklets were randomly administered to test-takers.

Table 2. Balanced Incomplete Block Design in the PISA 2009 Main Survey.

Booklet ID (number of persons)	Cluster (number of items)			
1 (445)	**M1 (9)**	R1	R3	**M3 (11)**
2 (445)	R1	S1	R4	R7
3 (447)	S1	R3	**M2 (12)**	S3
4 (452)	R3	R4	S2	R2
5 (451)	R4	**M2 (12)**	R5	**M1 (9)**
6 (438)	R5	R6	R7	R3
7 (452)	R6	**M3 (11)**	S3	R4
8 (452)	R2	**M1 (9)**	S1	R6
9 (448)	**M2 (12)**	S2	R6	R1
10 (441)	S2	R5	**M3 (11)**	S1
11 (450)	**M3 (11)**	R7	R2	**M2 (12)**
12 (449)	R7	S3	**M1 (9)**	S2
13 (461)	S3	R2	R1	R5

Source. Extracted from the Organization for Economic Co-Operation and Development (2012).
Note. "**M**" represents math, "R" represents reading, and "S" represents science. Bold values represent the structure of the PISA 2009 Taiwan math data set. PISA = Program for International Student Assessment.

The structure of the PISA 2009 Taiwan math data set (highlighted in bold in Table 2) was used for simulation in this study and the data were analyzed. There were 6,046 students and the ratio of boys to girls was nearly 1:1. Each item was administered to about 1,850 students, about 30% of the total sample. In other words, the missingness rate for each item was approximately 70%. This missingness was caused by design, so it could be considered as missing completely at random (Rubin, 1976) and could be ignored in maximum likelihood estimation (Mislevy & Wu, 1988). The authors followed this perspective to treat the designed missingness in this study.

Unlike most exams where the missingness is often very low (e.g, less than 5%), the BIB design in the PISA can have over 50% of missingness. In the PISA 2009 Taiwan math test, there were 4,035 participants included in the math data set, excluding those who did not receive math test items. However, each item had only about 1,800 test-takers and lead to a 56% missing rate. In general, the smaller the sample size, the lower the power of the DIF assessment. It would be overconfident to treat the result and power of a DIF assessment derived from the total sample size, which represents only 44% of the total responses. The missing rate might impair the performance of the DFTD and DSP on DIF assessment.

An Empirical Example

The PISA 2009 Taiwan math test data set was analyzed, which had 4,035 students (2,027 boys and 2,008 girls) responding to 32 dichotomous items in nine booklets. Due to the BIB design given in Table 2, 56% of responses were missing by design. Each math item was answered by about 1,800 students. Some students answered as few as nine math items, whereas some answered as many as 23 math items.

Before DIF assessment, the authors evaluated the model-data fit with the item infit and outfit mean square errors, whose values ranged from 0.87 to 1.13 with a mean of 1.00, indicating a satisfactory fit (Wright & Linacre, 1994). Next, the DFTD and DSP were adopted to assess gender DIF. It was found that the DFTD strategy yielded zero DIF items, whereas the DSP procedure yielded two DIF items (M496Q02 and M800Q01). Item M496Q02 had a difficulty of

−1.07 logits for boys and it favored girls by 0.26 logits, which was considered as a negligible DIF according to Educational Testing Service's classification (Zwick, Thayer, & Lewis, 1997). Item M800Q01 had a difficulty of −3.16 logits for boys and it favored boys by 0.56 logits, which was considered as a slight to moderate DIF. The mean ability of male respondents was higher than that of female respondents by over 0.4 logits (i.e., impact). The DIF magnitude of Item M496Q02 (0.26 logits), although statistically significant at the .05 nominal level, was not substantial and could be treated as practically insignificant (Wang, 2008). In short, the DFTD and DSP yielded very similar DIF assessment results, with the DSP procedure yielding two DIF items and the DFTD strategy yielding none. In addition, the ConQuest command, design matrix, and data file for this final Rasch DIF model estimation were provided as online supplements.

Simulation Study I

Design and Analysis

The authors conducted a series of simulations to compare the performance of the DFTD with DSP in gender DIF assessment. The data structure mimicked that of the empirical example that included 4,050 students (half of them were boys) and 32 items in nine booklets. The DFTD and DSP were compared under the following manipulated conditions:

1. Data types: complete and missing. In complete data, all students responded to all 32 items. In missing data, the BIB design in Table 2 was adopted.
2. Impact: 0 and 0.5. The reference group (i.e., boys) followed $N(0, 1)$, whereas the focal group (i.e., girls) followed either $N(0, 1)$ or $N(0.5, 1)$. In other words, the impact had two levels: 0 and 0.5 logits. The difficulty parameters of the 32 items for the boys group, shown in Table 3, were obtained from the empirical example.
3. Number of DIF items: 0, 6, and 13 items, which were similar to those found or adopted in empirical or simulation studies (Kopf et al., 2015; Lyons-Thomas, Sandilands, & Ercikan, 2014). The percentage of DIF items in the test was 0%, 19%, and 41%, respectively.
4. DIF magnitude: 0.5 and 1. The difficulty parameters of DIF items for the girl group were created by adding 0.5 or 1 logit to those difficulties for the boy group, as shown in Table 3. These two levels represented a moderate and large effect, respectively (Wang, 2004).

A total of 100 data sets (replications) were made under each condition and were analyzed with the ConQuest software (Adams, Wu, & Wilson, 2012), which incorporates marginal maximum likelihood estimation and the expectation–maximization algorithm for parameter estimation and observed Fisher's information for asymptotic standard error estimation. When the data were analyzed with the true model, the parameters were recovered pretty well, with bias values very close to zero and root mean square errors close to the estimated standard errors. The ICI method was used for locating DIF-free items in both the DFTD and DSP, but the DFTD used the CI method (without purification procedure) for DIF assessment, whereas the DSP used the AOI-P method. In ConQuest, the authors have to import corresponding design matrices manually for the ICI, CI, or AOI-P method. Dependent variables were the item-level Type I error rate and power rate of DIF detection. The Wald test with the .05 nominal level was used to assess DIF. Since the standard error estimation is critical to the Wald test, the authors adopted the ''SE=full'' option in ConQuest to avoid underestimation. However, the standard error estimation would be more challenging for complicated IRT models (such as two- or three-parameter logistic models), so the likelihood ratio test is recommended in those cases. The Type I error

Table 3. Item Parameters for the Boy Group in the Simulation Studies.

Item	Cluster	Item code in PISA 2009	Parameter
1	1	M033Q01	−1.069
2	1	M034Q01T	0.296
3	1	M155Q01	−0.272
4	1	M155Q04T	0.171
5	1	M411Q01	0.510
6	1	M411Q02	−0.001
7	1	M442Q02	0.480
8	1	M474Q01	−0.580
9	1	M803Q01T	1.089
10	2	M273Q01T	−0.239
11	2	M408Q01T	0.325
12	2	M420Q01T	0.141
13	2	M446Q01	−0.872
14	2	M446Q02	2.473
15	2	M447Q01	−1.350
16	2	M464Q01T	0.593
17	2	M559Q01	−1.468
18	2	M800Q01	−3.426
19	2	M828Q01	0.670
20	2	M828Q02	−0.008
21	2	M828Q03	1.513
22	3	M192Q01T	0.035
23	3	M406Q01	0.558
24	3	M406Q02	1.610
25	3	M423Q01	−2.332
26	3	M496Q01T	−0.390
27	3	M496Q02	−0.937
28	3	M564Q01	−0.119
29	3	M564Q02	0.016
30	3	M571Q01	0.914
31	3	M603Q01T	0.860
32	3	M603Q02T	0.806

Note. PISA = Program for International Student Assessment.

rate was computed as the number of times in 100 replications that a DIF-free item was mistakenly deemed as exhibiting DIF, and the power rate as the number of times that a DIF item was correctly deemed as exhibiting DIF. It was anticipated that the DSP procedure would outperform the DFTD strategy, especially where tests contained many DIF items and adopted the BIB design.

Results

The first step in both the DFTD and DSP was to search for a small set of DIF-free items to serve as anchors. Before discussing the Type I error and power rate of the DIF assessments, the accuracy rates using the ICI method in locating true DIF-free items as anchors should be outlined. The authors followed the common practice of identifying four items using the ICI method. Across the data types (two levels, complete and BIB), impacts (two levels, 0 and 0.5), number of DIF items (two levels, 6 and 13 items), and DIF magnitudes (two levels, 0.5 and 1), the accuracy rate (all four selected items were indeed DIF-free) was 100%, except under two conditions:

Table 4. Mean Type I Error Rates and Mean Power Rates in the DIF Assessment Using the DFTD Strategy and DSP Procedure.

Impact	DIF magnitude	Number of DIF items	Mean Type I error rate				Mean power rate			
			Complete		BIB		Complete		BIB	
			DFTD	DSP	DFTD	DSP	DFTD	DSP	DFTD	DSP
0	0.5	0	0.04	0.05	0.03	0.05				
		6	0.06	0.05	0.04	0.06	0.99	1.00	0.95	0.96
		13	0.13	0.05	0.11	0.05	0.99	0.99	0.85	0.96
	1	0	0.04	0.05	0.03	0.05				
		6	0.05	0.04	0.04	0.05	1.00	1.00	1.00	1.00
		13	0.12	0.05	0.11	0.05	1.00	1.00	1.00	1.00
0.5	0.5	0	0.04	0.04	0.03	0.04				
		6	0.06	0.05	0.04	0.05	0.97	1.00	0.92	0.97
		13	0.12	0.04	0.13	0.06	0.99	0.99	0.82	0.95
	1	0	0.04	0.05	0.03	0.05				
		6	0.06	0.05	0.04	0.05	1.00	1.00	0.97	1.00
		13	0.13	0.06	0.12	0.05	1.00	1.00	0.99	0.99

Note. Mean Type I error rates that are beyond (2.5%, 7.5%) are considered inappropriate and underlined; their corresponding power rates are underlined as well. DIF = differential item functioning; DFTD = DIF-free-then-DIF strategy; DSP = dual-scale purification; BIB = balanced incomplete block.

(a) data type = BIB, impact = 0, number of DIF items = 13, DIF magnitude = 0.5; and (b) data type = BIB, impact = 0.5, number of DIF items = 13, DIF magnitude = 0.5. Under these two conditions, the accuracy rate was 99%. Consistent with the literature (Shih & Wang, 2009; Wang et al., 2012), the ICI method performed almost perfectly in identifying a small set of DIF-free items.

Treating the four selected items as anchors, the second step in the DFTD strategy was to assess DIF in the other items. The DSP procedure proceeded as previously stated. Due to space constraints, the authors have not reported the Type I error rates and power rates for individual items (these are available on request); instead, they have shown the mean Type I error rate across DIF-free items and the mean power rate across DIF items in Table 4. The standard deviation, maximum, and minimum values of the Type I error rates and power rates over replications in each condition are listed in the online supplements. A mean Type I error rate beyond the range of 0.025 to 0.075 was considered inappropriate, which made its corresponding power rate meaningless. It can be seen in Table 4 that the DFTD strategy yielded inflated mean Type I error rates under eight of the 24 conditions, whereas the DSP procedure always yielded well-controlled mean Type I error rates. When there were as many as 13 DIF items in the tests (i.e., 41% DIF items in the test), the DFTD strategy slightly lost control of Type I error rates, ranging from 0.11 to 0.13.

Next, the mean power rates are considered. In complete data conditions, the DFTD and DSP both yielded almost the same and perfect power rates. In the BIB design, the DSP procedure yielded higher power rates than those of the DFTD strategy, especially when DIF magnitude = 0.5 and the number of DIF items = 13. There might be two reasons for this. First, the DFTD strategy might incorrectly select DIF items as anchors under these conditions (e.g., the 99% accuracy rate). Second, the number of anchors for the DFTD was only four. The DSP procedure yielded similar power rates in the complete data and the BIB design, whereas the DFTD

strategy yielded higher power rates in the complete data than in the BIB design, suggesting that only the DSP procedure could maintain high power rates when there was many missing data.

In this study, the numbers of persons answering an item were around 1,800 in the BIB design. It was interesting to investigate whether the DSP procedure would yield similar Type I error rates and power rates when all items were responded by the same 1,800 students (the complete data design). As a demonstration, the authors simulated a complete data set with 1,800 persons together with impact = 0, DIF magnitude = 0.5, number of DIF items = 13, and DIF strategy = DSP. The resulting mean Type I error rate was 0.04 and the mean power rate was 0.97, which were very similar to those found when there were 4,050 students with the BIB design (see Table 4).

One might also question how the DSP procedure would perform when the missing rate was higher. As a demonstration, the authors removed 10%, 20%, and 30% of item responses randomly under the condition of 4,050 students, BIB design, impact = 0, DIF magnitude = 0.5, number of DIF items = 13, and DIF strategy = DSP. The resulting mean Type I error rates were 0.05, 0.05, and 0.06, and the mean power rates were 0.94, 0.91, and 0.90 for the 10%, 20%, and 30% missing rates, respectively. From these two additional brief simulations, it could be concluded that Type I error rates were not affected by sample sizes, but power rates were largely determined by the number of persons who actually responded to items. The results were consistent with the DIF literature (e.g., Kopf et al., 2015; Paek & Wilson, 2011; Wang et al., 2012).

In this study, the authors set the impact at 0 or 0.5 and did not simulate very large impacts. In reality, the impact can be larger than 0.5. The DSP procedure should have performed similarly if larger impacts had been simulated. To verify, the authors conducted another brief simulation study where impact = 1, together with the BIB design, DIF magnitude = 0.5, number of DIF items = 13, and DIF strategy = DSP. It was found that the mean Type I error rate and power rate were 0.08 and 0.91, respectively, which were very similar to but slightly worse than those of 0.06 and 0.95, respectively, under the impact = 0.5 condition. The high similarity was expected because the Rasch DIF model had portioned out the effect of impacts from the DIF effect.

Simulation Study 2

In the previous simulation study, the DSP outperformed the DFTD on mean power rates. One of the reasons is the number of anchor items. The DFTD adapts four presumed DIF-free items as anchors, whereas the DSP iteratively includes DIF-free items in the anchor. The mean of the number of anchor items for the last purification run of the DSP was 19 when the number of DIF items was 13, and was 25 when the number of DIF items was six, which are a lot more than the four items used in DFTD. It has been found that the more the clean items as anchors (containing exclusively DIF-free items), the higher the power would be (Wang & Yeh, 2003). In this simulation study, the authors increased the number of anchored items from four to seven, 10, 13, 16, and 19 for the DFTD strategy. However, they focused only on one of the two conditions where the DFTD performed substantially worse than the DSP (i.e., data type = BIB, impact = 0, number of DIF items = 13, and DIF magnitude = 0.5).

As shown in Table 5, the greater the number of items selected as anchors, the lower the accuracy rate of locating true DIF-free items as anchors. The accuracy rate was 99% when four items were selected, whereas it became only 78% when 19 items were selected. The mean Type I error rates ranged from 0.11 to 0.15. The mean power rate increased from 0.85 with four items to 0.90 with 10 items but decreased to 0.67 with 19 items. Overall, an anchor length of 10 items (out of 32 items) would have the highest power rate and a rather less inappropriate Type I error rate. However, the performance of the DFTD with a 10-item anchor was still poorer than the performance of the DSP under the same conditions (i.e., the mean Type I error rate = 0.06 and the mean power rate = 0.96, as denoted in Table 4).

Table 5. Anchor Lengths, Accuracy Rates, Mean Type I Error Rates, and Mean Power Rates in the DIF Assessment Using the DFTD Strategies.

Anchor length	Accuracy rate	Mean Type I error rate	Mean power rate
4	.99	0.11	0.85
7	.97	0.12	0.88
10	.95	0.11	0.90
13	.90	0.12	0.87
16	.83	0.14	0.78
19	.78	0.15	0.67

Note. DIF = differential item functioning; DFTD = DIF-free-then-DIF strategy.

Conclusion and Discussion

The importance of a clean common-scale DIF assessment is emphasized, and scale purification procedures are widely advocated. Unfortunately, when tests consist of many DIF items, the scale is too contaminated by the inclusion of these DIF items to be purified, and thus the DIF assessment is jeopardized. The DFTD strategy was proposed to resolve this problem. In past studies, it has been found that the DFTD strategy, although outperforming traditional-scale purification methods (e.g., AOI-P) when tests consist of a high percentage of DIF items, fails to control Type I error rates under such conditions. Although the DFTD strategy has two steps (the anchor-item-searching step and the DIF-testing step), only the anchor-item-searching step has a scale purification procedure. It might be desirable to add another scale purification procedure to the DIF-testing step, which was implemented in this study and was referred to as the DSP procedure. It was hoped that adding another scale purification procedure would improve DIF assessment.

The DIF assessment methods have not been thoroughly investigated when data consist of a high percentage of missingness, such as when the BIB design is implemented in large-scale programs. The authors thus investigated the DFTD and DSP under such conditions using the design of the PISA 2009 Taiwan math test data as an example. These simulation studies demonstrate the superiority of the DSP procedure over the DFTD strategy when the BIB design is adopted. When tests consisted of a high percentage of DIF items, the DFTD strategy yielded slighted inflated Type I error rates and deflated power rates. The DSP procedure yielded well-controlled Type I error rates and high power rates even under such conditions. Compared with the complete data condition, the BIB design worsened the power of the DFTD strategy more seriously than that of the DSP procedure. Increasing the number of anchored items in the DFTD strategy from four to 10 items (in a 32-item test) yielded little improvement on the power. It seems that the DSP procedure is especially promising in DIF assessment when tests consist of many DIF items or when there is a high percentage of missingness. In practice, one will never be exactly sure about the percentage of DIF items in a test, so the DSP procedure is preferred.

It may be very tedious for practitioners to implement the DFTD strategy, not to mention the DSP procedure. The authors are developing freeware in R for the DFTD strategy and DSP procedure to facilitate their utility. The AOI-P method in the DSP procedure requires several iterations of purification, and the iteration stops when the DIF results of two consecutive runs are the same (i.e., converged). However, the stopping rule may not be satisfied from time to time. As per the authors' experience, the chance of nonconvergence is as low as 0.67%. In practice, one may set a maximum number of iterations, such as 15 iterations, to force convergence.

The DSP procedure was developed within an IRT framework and was investigated under very limited conditions, such as equal sample sizes between groups or small impacts. In general, if the sample size of the reference group is fixed, then, the larger the sample size of the focal group, the higher the power of DIF detection. On the contrary, if the total sample size of the reference and focal groups is fixed, then, the more similar the sample sizes between groups (e.g., equal sample size as an extreme case), the higher the power of DIF detection (Awuor, 2008; Cuevas & Cervantes, 2012; Paek & Guo, 2011). The impact was set at 0 and 0.5 in this study. The authors conducted a brief simulation where impact = 1 and observed a slight deterioration in the Type I error rate and power rate. More simulations are needed to investigate how the DSP procedure performs when sample sizes are very different between groups and/or when impact = 1 or more extreme values.

The effects of missingness rates and linking designs on the DSP procedure were not investigated thoroughly because the authors aimed to mimic the PISA situation where the missingness rate was not serious and the BIB design was adopted. The Rasch DIF model, which assumes a common slope across items and homogeneous variances between groups, was used in this study to demonstrate the advantages of the DSP. Future studies can be conducted to examine the performance of the DSP when missingness rates are serious, other linking designs are adopted, different groups have different variances, or different IRT models or non-IRT-based DIF methods (e.g., the MH method) are adopted.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Supplemental Material

Supplementary material is available for this article online.

References

Adams, R. J., Wu, M. L., & Wilson, M. R. (2012). ACER ConQuest: Generalised item response modelling software (Version 3) [Computer software]. Camberwell, Victoria: Australian Council for Educational Research.

Allen, N. L., & Donoghue, J. R. (1996). Applying the Mantel-Haenszel procedure to complex samples of items. *Journal of Educational Measurement*, *33*, 231-251.

Awuor, R. A. (2008). *Effect of unequal sample sizes on the power of DIF detection: An IRT-based Monte Carlo study with SIBTEST and Mantel-Haenszel procedures* (Unpublished doctoral dissertation). Virginia Polytechnic Institute and State University, Blacksburg, VA.

Candell, G. L., & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement*, *12*, 253-260.

Chen, C.-T., Wang, W.-C., & Shih, C.-L. (2012, July). *Effect of scale purification on the assessment of differential rater functioning*. Paper presented at the 77th Annual Meeting of the Psychometric Society, Lincoln, NE.

Chen, J.-H., Chen, C.-T., & Shih, C.-L. (2014). Improving the control of Type I error rate in assessing differential item functioning for hierarchical generalized linear model when impact is presented. *Applied Psychological Measurement*, *38*, 18-36.

Cuevas, M., & Cervantes, V. H. (2012). Differential item functioning detection with logistic regression. *Mathematics and Social Sciences*, *199*, 45-59.

Finch, H. (2011). The use of multiple imputation for missing data in uniform DIF analysis: Power and Type I error rates. *Applied Measurement in Education*, *24*, 281-301.

Goodman, J. T., Willse, J. T., Allen, N. L., & Klaric, J. S. (2011). Identification of differential item functioning in assessment booklet designs with structurally missing data. *Educational and Psychological Measurement*, *71*, 80-94.

Grisay, A., Gonzalez, E., & Monseur, C. (2009). Equivalence of item difficulties across national versions of the PIRLS and PISA reading assessments. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments*, *2*, 63-83.

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

Kopf, J., Zeileis, A., & Strobl, C. (2015). Anchor selection strategy for DIF analysis: Review, assessment, and new approaches. *Educational and Psychological Measurement*, *75*, 22-56.

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

Lyons-Thomas, J., Sandilands, D., & Ercikan, K. (2014). Gender differential item functioning in mathematics in four international jurisdictions. *Education and Science*, *39*, 20-32.

Mislevy, R. J., & Wu, P. K. (1988). *Inferring examinee ability when some item responses are missing* (ERIC Document Reproduction Service No. ED 395 017). Princeton, NJ: Education Testing Service

Oort, F. J. (1998). Simulation study of item bias detection with restricted factor analysis. *Structural Equation Modeling*, *5*, 107-124.

Organization for Economic Co-Operation and Development. (2002). *PISA 2000 technical report*. Paris, France: Author.

Organization for Economic Co-Operation and Development. (2012). *PISA 2009 technical report*. Paris, France: Author.

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

Paek, I., & Wilson, M. (2011). Formulating the Rasch differential item functioning model under the marginal maximum likelihood estimation context and its comparison with Mantel–Haenszel procedure in short test and small sample conditions. *Educational and Psychological Measurement*, *71*, 1023-1046.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests* (Expanded ed.). Chicago, IL: The University of Chicago Press.

Robitzsch, A., & Rupp, A. A. (2009). Impact of missing data on the detection of differential item functioning. *Educational and Psychological Measurement*, *69*, 18-34.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581-592.

Sandilands, D. A. (2014). *Accuracy of differential item functioning detection methods in structurally missing data due to booklet design* (Unpublished doctoral thesis). The University of British Columbia, Vancouver, Canada.

Shealy, R. T., & Stout, W. F. (1993). A model-biased 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.

Shih, C. L., & Wang, W. C. (2009). Differential item functioning detection using the multiple indicators, multiple causes method with a pure short anchor. *Applied Psychological Measurement*, *3*, 184-199.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*, 361-370.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 147-169). Hillsdale, NJ: Lawrence Erlbaum.

Wang, W. C. (2004). Effects of anchor item methods on the detection of differential item functioning within the family of Rasch models. *The Journal of Experimental Education*, *72*, 221-261.

Wang, W. C. (2008). Assessment of differential item functioning. *Journal of Applied Measurement*, *9*, 387-408.

Wang, W. C., Shih, C. L., & Sun, G. W. (2012). The DIF-free-then-DIF strategy for the assessment of differential item functioning. *Educational and Psychological Measurement*, *72*, 687-708.

Wang, W. C., & Yeh, Y. L. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement*, *27*, 479-498.

Woods, C. M. (2009). Empirical selection of anchors for tests of differential item functioning. *Applied Psychological Measurement*, *33*, 42-57.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement: Transactions of the Rasch Measurement SIG*, *8*, 370.

Zwick, R., Thayer, D. T., & Lewis, C. (1997). *An investigation of the validity of an empirical Bayes approach to Mantel-Haenszel DIF analysis* (ETS Research Report No. 97-21). Princeton, NJ: Educational Testing Service.