Equated Pooled Booklet Method in DIF Testing

Ying Cheng, Peihua Chen, Jiahe Qian and Hua-Hua Chang

Applied Psychological Measurement 2013 37: 276 originally published online 28 January 2013
DOI: 10.1177/0146621612471889

Article

Equated Pooled Booklet Method in DIF Testing

Applied Psychological Measurement
37(4) 276–288

DOI: 10.1177/0146621612471889
apm.sagepub.com

Ying Cheng[1], Peihua Chen[2], Jiahe Qian[3], and Hua-Hua Chang[4]

Abstract

Differential item functioning (DIF) analysis is an important step in the data analysis of large-scale testing programs. Nowadays, many such programs endorse matrix sampling designs to reduce the load on examinees, such as the balanced incomplete block (BIB) design. These designs pose challenges to the traditional DIF analysis methods. For example, as difficulty levels often vary across booklets, examinees with same booklet scores may be disparate in ability. Consequently, DIF procedures based on matching total scores at the booklet level may cause misplacement of examinees and inflation in measurement errors. Therefore, modification to traditional DIF procedures to better accommodate the BIB design becomes important. This article introduces modification of current simultaneous item bias test (SIBTEST) procedure for the DIF analysis method when multiple booklets are used. More specifically, examinees will be pooled across booklets, and the matching will be based on transformed booklet scores after common block equating/linking. Simulations are conducted to compare the performance of this new method, the *equated pooled booklet* method against that of the current *pooled booklet* method, in terms of both Type I error control and power. Four factors are considered in the simulation—the DIF effect size, item difficulty, impact, and the length of common block. Results show that the *equated pooled booklet* method in general improves power while keeping Type I error under control. The advantage of the new method is the most pronounced when the traditional method struggles, for example, when the item is difficult or there is impact.

Keywords

SIBTEST, DIF, equating, booklet design, polySIBTEST

Introduction and Background

Differential item functioning, or DIF, is a very important step in data analysis of many large-scale testing programs, such as the Programme for International Student Assessment (PISA;

[1]University of Notre Dame, IN, USA
[2]National Chiao Tung University, Hsinchu City, Taiwan
[3]Educational Testing Service, Princeton, NJ, USA
[4]University of Illinois at Urbana–Champaign, USA

Corresponding author:
Ying Cheng, University of Notre Dame, 118 Haggar Hall, Notre Dame, IN 46556, USA.
Email: ycheng4@nd.edu

Yildirim & Berberoglu, 2009), Trends in International Mathematics and Science Study (TIMSS; Robitaille & Beaton, 2002), and the National Assessment of Educational Progress (NAEP, see http://nces.ed.gov/nationsreportcard/tdw/analysis/scaling_checks_dif_proced.asp). The purpose of DIF analysis is to identify items that function differently for examinees of the same underlying ability from different subgroups. More specifically, suppose that two groups of interest take a test, one group called the focal group (F) and the other called the reference group (R). DIF studies compare the performance of the two groups to assess whether items are potentially biased against the examinees from one of the groups.

As an example, the 1998 NAEP assessed nearly 448,000 students in the national and state samples. DIF analyses were conducted for items in reading, writing, and civics. The analyses involved three reference group/focal group comparisons: male/female, White/Black, and White/Hispanic (Allen, Carlson, & Donoghue, 2001). The DIF detection procedures used in NAEP data analysis include the Mantel–Haenszel (M-H) procedure (Holland & Thayer, 1988; Mantel & Haenszel, 1959), standardization procedure (Dorans & Kulick, 1986; Dorans, Schmitt, & Bleistein, 1992), and the simultaneous item bias test (SIBTEST) procedure (Shealy & Stout, 1993). Other DIF detection methods used in testing industry include the logistic regression procedures and hierarchical linear model procedure (Swaminathan & Rogers, 1990; Swanson, Clauser, Case, Nungester, & Feathermean, 2002). This article focuses on the SIBTEST procedure.

A Review of the SIBTEST Procedure

The SIBTEST procedure was originally developed by Shealy and Stout (1993). It adopted a null DIF definition by requiring that the regression of item score on the latent trait be identical for the groups under study.

Definition 1. Let $E_R[Y|\theta]$ and $E_F[Y|\theta]$ denote the expected value of Y on θ for reference and focal groups, respectively. An item does NOT exhibit DIF if

$$E_R[Y|\theta] = E_F[Y|\theta] \tag{1}$$

for all values of θ, where Y and θ represent the observed response and the latent trait, respectively.

Hence, local DIF at a given θ level can be measured by

$$B(\theta) \equiv E_R[Y|\theta] - E_F[Y|\theta] \tag{2}$$

Shealy and Stout (1993) proposed a global index of DIF for the dichotomous case

$$\beta = \int B(\theta) f_F(\theta) \mathrm{d}\theta, \tag{3}$$

where $f_F(\theta)$ is the density function of θ in the focal group. SIBTEST performs DIF detection by testing

$$H_0 : \beta = 0 \text{ versus } H_1 : \beta \neq 0. \tag{4}$$

According to Chang, Mazzeo, and Roussos (1996), Definition 1 is equivalent to Definition 2.

Definition 2. Let $E_R[Y|t]$ and $E_F[Y|t]$ denote the expected value of Y on true score t for reference and focal groups, respectively. An item does not exhibit DIF if

$$E_R[Y|t] = E_F[Y|t] \tag{5}$$

for all values of t.

Analogously to Equation 2, the local DIF at t can be expressed as

$$D(t) = E_R[Y|t] - E_F[Y|t]. \tag{6}$$

Thus, the global DIF can also be defined as

$$\beta = \int D(t) f_F(t) \mathrm{d}t. \tag{7}$$

Intuitively, one would expect that DIF could be estimated locally by the values of

$$d_k = \bar{Y}_{Rk} - \bar{Y}_{Fk}, \ k = 0, 1, 2, \ldots, n_H, \tag{8}$$

where $\bar{Y}_{Rk} - \bar{Y}_{Fk}$ is the group difference on the studied item among examinees with the same observed matching score k, and n_H is the highest possible matching score. If examinees with the same matching test *observed* score have the same *true* score, then Equation 7 is also approximately the difference in item scores at the same true score. If the studied item does not have observed-score DIF, it is expected that $d_k \approx 0$. Thus, a suggested statistic to estimate the global DIF β for the special case where the reference and focal groups have the same ability distribution is

$$\hat{\beta} = \sum_{k=1}^{n_H} p_k d_k, \tag{9}$$

where p_k is the proportion of examinees with the observed matching score k. According to Shealy and Stout (1993), a test statistic can be defined by

$$B = \frac{\hat{\beta}}{\hat{\sigma}(\hat{\beta})} \tilde{N}(0, 1), \tag{10}$$

where $\hat{\sigma}(\hat{\beta})$ is the standard deviation of $\hat{\beta}$. The procedure neither requires nor uses IRT ability or item parameter estimates for its calculation.

Actually, SIBTEST does true-score-matched DIF estimation by performing linear regressions of $T_{F,k}$ and $T_{R,k}$, where $T_{g,k}$ denotes the regression of the true score, given the observed score k in group g, $g = F$ or R. This transformation is referred to as the *regression correction* by Shealy and Stout (1993). Note that with the same matching *observed* score k, $T_{F,k}$ and $T_{R,k}$ can differ when the focal and reference groups have different means and reliability, and may not be whole numbers. Their average, T_k (again, may not be a whole number), then serves as the point where two groups' average score on the interested item are matched. In other words, $\bar{Y}_{Rk}$ and $\bar{Y}_{Fk}$ become $\bar{Y}_{R,T_k}$ and $\bar{Y}_{F,T_k}$, respectively. For simplicity's sake, $\bar{Y}_{g,T_k}$ will be denoted as $\bar{Y}^*_{gk}$, where $g = F$ or R, and they replace the $\bar{Y}_{gk}$ in Equation 8. For details of the correction, see Shealy and Stout (1993), and Bolt (1996).

The original SIBTEST developed by Shealy and Stout (1993) was based on dichotomous items. According to the theoretical result of Chang and Mazzeo (1994), Equations 1, 2, and 3, or Equations 5, 6, and 7, form an appropriate basis for testing DIF in both dichotomous and polytomous items. A polytomous procedure was developed by making relatively minor modifications to the dichotomous SIBTEST procedure (Chang et al., 1996). The modified procedure, named polySIBTEST, includes dichotomous item scoring as a special case. In using polySIBTEST, the studied item can be either dichotomous or polytomous, and the matching test can consist of a mixture of both types of items.

Table 1. Example of the BIB Booklet Design.

	Booklet 1	Booklet 2	Booklet 3	Booklet 4	Booklet 5	Booklet 6	Booklet 7
Block Position 1	A	D	F	B	E	G	C
Block Position 2	B	A	G	E	C	D	F
Block Position 3	C	E	A	F	G	B	D

Note: BIB = balanced incomplete block.

Balanced Incomplete Block (BIB) Design

It should be noted that the three procedures, M-H, standardization, and (poly)SIBTEST, were all originally developed for use with traditional test format, that is, all examinees receiving an identical test. However, in modern large-scale testing programs such as NAEP (Zwick, 1987) and PISA, different test booklets are used to limit test time, which means each examinee only needs to take a portion of the items rather than all of them (Frey, Hartig, & Rupp, 2009). This is known in testing as *multiple matrix sampling*, a term that arises from ''giving samples of items to samples of examinees'' (Gonzalez & Rutkowski, 2010; Mislevy, Johnson, & Muraki, 1992). It is therefore important to address DIF detection in tests with multiple booklets.

Among various matrix sampling booklet designs, Lord (1965) suggested using the BIB design in large-scale assessment. For each assessment subject, the items are grouped into several separately timed groups, termed *blocks*. The BIB design ensures that every item block appears an equal number of times in all block positions. Currently, the BIB design is used in the construction of the NAEP cognitive test booklets. More specifically, item blocks are then combined into several test booklets consisting of three blocks each. The design is organized so that each block appears in each position (first, second, or third) within a booklet, and each pair of blocks appears together the same number of times. See Table 1 for an example of the BIB design. Such design was used in the 1990 NAEP Math Assessment (Allen & Donoghue, 1996) where seven blocks spiraled through seven booklets, with each booklet containing three blocks. For instance, as shown in Table 1, the first booklet consists of Blocks A, B, and C, in that order. The same design was also explained in Frey et al. (2009). Another example can be found at http://nces.ed.gov/nationsreportcard/tdw/instruments/cog_spiral.asp, where five blocks spiraled through 10 booklets. Each booklet is administered to a random sample of students.

Such complex sampling of items results in sparse responses for individual items. This raises questions about the appropriateness of traditional approaches in forming matching variables for these popular DIF procedures. This article discusses an approach to deal DIF detection with multiple booklets. Note that even though BIB design was used to illustrate the method, the approach itself can be generally applied to any booklet design with minimal adaptation when the booklets share common items.

PolySIBTEST Procedure With Booklet Design

As described previously, in the BIB design, one studied item is contained in three different booklets. Assume the studied item is in Block A from the complex sampling design portrayed in Table 1. The central problem is that for this specific studied item, there are three different $\hat{\beta}$s that are defined in Equation 9 for Booklet 1, 2, and 3 (denoted as $\hat{\beta}_1, \hat{\beta}_2,$ and $\hat{\beta}_3$, respectively). Now they need to be combined as a single index. The current procedure takes a weighted average of them. More specifically, the $\hat{\beta}$ statistic, now denoted as $\hat{\beta}_{old}$, can be expressed by the following equation

$$\hat{\beta}_{old} = W_1\hat{\beta}_1 + W_2\hat{\beta}_2 + W_3\hat{\beta}_3, \quad (11)$$

where W_1, W_2, and W_3 are sample-size-determined weights, and $W_1 + W_2 + W_3 = 1$. Analogous to Equation 10, the corresponding test statistic, B, now called B_{old}, is obtained as follows:

$$B_{old} = \frac{\hat{\beta}_{old}}{\sqrt{W_1^2\hat{\sigma}^2(\hat{\beta}_1) + W_2^2\hat{\sigma}^2(\hat{\beta}_2) + W_3^2\hat{\sigma}^2(\hat{\beta}_3)}}. \quad (12)$$

Potential Problems With the Current Procedure

The procedure just described is similar to the booklet-level matching (Allen & Donoghue, 1996) method, which means each booklet yields a different statistic for the same item in the common block. In other words, an item can have multiple measures of DIF. Another potential drawback of booklet-level matching is that the number of examinees receiving the same booklet is much smaller than the number of examinees receiving a common block of items. Therefore, the individual DIF statistics calculated at the booklet level will be subject to greater sampling variability than will those computed with block-level matching. Hence, the standard error associated with the DIF statistic tends to be larger, and the test based on the DIF statistics calculated from the booklet-level matching would be less powerful than that based on pooling all examinees across booklets.

In addition, if the difficulty levels vary across booklets, examinees with the same total booklet score may be disparate in ability. Therefore, matching based on total booklet score may cause misplacement and increase measurement errors. In fact, difficulty levels do vary across booklets in reality. For example, according to Qian Steven, Lois, and Liang (2001), in NAEP Grade 8 public schools sample of the 1998 state reading assessment, the average block difficulty level varied from 0.41 to 0.69. (See Table 17-3 in Qian et al., 2001.) Thus, alternative matching variables are desirable.

Research Design

The current SIBTEST procedure is remedied based on total-booklet-score matching (called ''current procedure'' thereafter) in two ways:

1. Equate the booklet score using common block equating/linking.
2. Use pooled booklet matching for DIF detection based on the transformed booklet score.

The initial step is fairly straightforward. The pooled booklet matching was first proposed in Allen and Donoghue (1996), and was used along with the M-H DIF detection. For a design with m booklets, instead of pooling m DIF statistics, each from a booklet, as shown earlier, they pooled all the examinees who responded to the same item of interest across m booklets, and computed the DIF statistic thereupon. The booklet scores are equated before pooling the examinees. The equated scores are rounded because the valid booklet scores are integers. After equating, the total scores from different booklets become more comparable. Hence, the misplacement in matching will be reduced. By using pooled booklet matching, the sample size of each score category will increase. Thus, the power of the test can be enhanced.

Table 2. Three Booklets Anchored by a Common Block.

Book 1	A	B	C
Book 2	D	A	E
Book 3	F	G	A

New Method

For simplicity of illustration, it is assumed that (a) the studied item is contained in the common Block A and (b) scores from each of the three booklets, Booklets 1, 2, and 3, all containing Block A, are used to form the matching variable. Similar assumptions have been made in the Allen and Donoghue (1996) study. Because Block A is present in all three booklets, it is said that these booklets are ''anchored'' by a common-item Block A (see Table 2).

To make scores comparable across the three booklets, one can perform certain score transformations that are based on the block of common items (Block A). This is well connected with the *common-item nonequivalent groups equating.* Specifically, the equating method implemented is Tucker's procedure (Kolen & Brennan, 1995). After equating, the total scores from the three booklets will be transformed to a common scale.

Each booklet is taken by different examinees and therefore there is a reference group and a focal group for each booklet. However, as the scores from different booklets are now transformed to a common scale, a large reference group that contains the three booklet-level reference groups and a large focal group that contains the three booklet-level focal groups are formed for the common items. Statistics can then be calculated based on the two large groups. Note that the computation of the test statistic B after pooling the booklets still follows Equations 9 and 10. Instead of pooling three individual $\hat{\beta}$s and their standard errors, the new method computes one single $\hat{\beta}$ and its standard error based on a pooled sample.

Simulation Studies

It is speculated that the new method will make more accurate DIF detection than the current method. However, this is too simple an answer to various situations encountered in DIF detection. First of all, the length of the common block can have significant impact on equating, and there is no uniform length of common block in booklet design. Take the 1992 National Mathematics Assessment in NAEP (Jenkins & Kulick, 1994) as an example, for Grade 8, the shortest block consisted of 9 items and the longest block consisted of 21 multiple-choice items. Usually, the longer the common block, the more accurate the equating. Therefore, the length of the booklet should be considered.

Second, the reference and focal groups may share the same latent trait distribution, or they may differ in this respect. The former is referred to as an impact-free condition. Whether impact is present has long been known to be influential on DIF detection. Some DIF detection procedures are effective under impact-free condition but lose power when impact is present. Both conditions are included in the simulation study.

Third, the severity of DIF, or effect size, will have an impact on DIF detection. Besides, the characteristics of the studied item, for instance, the difficulty level, might affect DIF detection, too.

Hence, this simulation study models (a) two levels of common block length—the short common block contains 10 items while the long one contains 20; (b) two levels of impact—one is impact free where the θs of both the reference group and the focal group are sampled from the

Table 3. Three Studied Items' Parameters.

Item	*a*	*b*	*c*
1	0.86	−1.63	0.19
2	0.68	0.01	0.20
3	0.67	1.52	0.14

standard normal distribution $N(0,1)$, and the impact condition is produced by sampling the reference group from $N(0,1)$ while sampling the focal group from $N(-1, 1)$; (c) three DIF levels—0.25, 0.5, and 1, featuring small, medium, and large effect sizes; and (d) three levels of item difficulty—for this purpose, three different studied items are considered. See Table 3 for parameters of the three items. Other details of the simulation are introduced as follows.

Test Structure. The data sets were generated by mimicking the booklet design described in Table 2. Three booklets are constructed, each consisting of three equal-length blocks. A common block of 10 or 20 items (Block A) appear in all three booklets. Among these 10 or 20 items, 3 items are studied for DIF. Consequently, the booklet length is either 30 or 60. The real item parameters released in the 1992 NAEP Technical Report (Johnson & Carlson, 1994) were used in conducting the simulation.

Item Response Generation. Ability (θ) values for 450 examinees for each of the reference and focal groups are randomly generated by sampling from the corresponding distribution according to the impact situation described earlier. An equal number of examinees (i.e., 150) will take each booklet. Item responses are generated based on the three-parameter logistic model.

Type I Error Rates Study. Type I error occurs when an item with no DIF is flagged as a DIF item. A no-DIF situation is created when the difficulty parameter of the studied item is kept the same across the focal and reference groups. The simulation is run with 1,000 replications. After each simulation, the item is either flagged as a DIF item (Type I error) or not flagged. The number of times the studied item is erroneously flagged is thus obtained. Dividing this number by 1,000 will give the Type I error rate.

Power Study. Power denotes the probability that an item with DIF being detected. In this study, a DIF situation is created by manipulating the difficulty parameter of the studied item. In DIF literature, this is known as parallel DIF (Hanson, 1998). There are other types of DIF, for example, directional DIF where the item response functions (IRFs) among focal versus reference group of examinees can cross. Parallel DIF is focused to keep the study manageable and to facilitate interpretation of findings. Mild DIF can be introduced by adding a small shift to the difficulty parameter *b* and applying it to the focal group. This makes the item ''appear'' more difficult for the focal group. Three levels of DIF are manipulated in this study by adding 0.25, 0.5, and 1.0, respectively, to the *b* parameter of the studied item. The simulation is replicated 1,000 times. After each replication, the item is either flagged as a DIF item (power) or not. The number of times the studied item is correctly flagged is thus obtained. Dividing the number by 1,000 is power.

Results

Tables 4 through 7 summarize the simulation results for the short common block situation (10 items) while Tables 8 through 11 do so for the long common block situation (20 items). Among them, Tables 4, 5, 8, and 9 represent the impact-free condition, and the others contain results when an impact is present.

Table 4. Type I Error Rates for Short Common Block (10 Items), Impact-Free Condition.

Item	New	Current
1	.07	.06
2	.06	.06
3	.07	.07
Average	.07	.06

Table 5. Power for Short Common Block (10 Items), Impact-Free Condition.

Item	DIF level	New	Current	Difference
1	Low	0.32	0.31	0.01
	Medium	0.84	0.83	0.01
	High	1.00	1.00	0.00
2	Low	0.30	0.29	0.01
	Medium	0.81	0.81	0.00
	High	1.00	1.00	0.00
3	Low	0.19	0.17	0.02
	Medium	0.47	0.42	0.05
	High	0.91	0.86	0.05
Average		0.65	0.63	0.02

Note: DIF = differential item functioning.

Table 6. Type I Error Rates for Short Common Block (10 Items), Impact Condition.

Item	New	Current
1	.18	.13
2	.10	.09
3	.08	.08
Average	.12	.10

Table 7. Power for the Short Common Block (10 Items), Impact Condition.

Item	DIF level	New	Current	Difference
1	Low	0.65	0.39	**0.26**
	Medium	0.95	0.82	**0.13**
	High	1.00	1.00	0.00
2	Low	0.32	0.21	**0.11**
	Medium	0.72	0.54	**0.17**
	High	0.99	0.95	0.04
3	Low	0.15	0.11	0.04
	Medium	0.30	0.19	**0.11**
	High	0.61	0.45	**0.17**
Average		0.63	0.52	0.11

Note: DIF = differential item functioning.

Table 8. Type I Error Rates of Long Common Block (20 Items), Impact-Free Condition.

Item	New	Current
1	.06	.05
2	.05	.05
3	.05	.04
Average	.06	.05

Table 9. Power for Long Common Block (20 Items), Impact-Free Condition.

Item	DIF level	New	Current	Difference
1	Low	0.30	0.26	0.04
	Medium	0.86	0.78	0.08
	High	1.00	1.00	0.00
2	Low	0.30	0.27	0.03
	Medium	0.80	0.70	**0.10**
	High	1.00	1.00	0.00
3	Low	0.16	0.11	0.05
	Medium	0.46	0.35	**0.10**
	High	0.90	0.80	**0.10**
Average		0.64	0.59	0.05

Note: DIF = differential item functioning.

Table 10. Type I Error Rates for Long Common Block (20 Items), Impact Condition.

Items	New	Current
1	.09	.11
2	.06	.07
3	.06	.06
Average	.07	.08

Table 4 shows that when the common block contains 10 items and there is no impact, both the new method and the current procedure control Type I error rates very well. The resulting Type I error rates are very close to the nominal level (i.e., 5%). The difference between the new versus current methods is negligibly small.

Table 5 demonstrates that for the short common block and impact-free condition, the new method produces slightly *higher* (on average about 2% higher; see the ''Difference'' column) power than the current method. Within each item, regardless of the DIF detection procedure, it is shown that the higher the DIF level (or effect size), the higher the power, which is expected. For instance, for the first item under low DIF, the power is about 0.30, whereas under high DIF, the power increases to 1.00. Across items, the more difficult the item is, the lower the power. For instance, under low DIF, for the easiest item (Item 1), the power under new method is 0.32; for the item with medium difficulty (Item 2), the power drops to 0.30; for the most difficult item (Item 3), the power further drops to 0.19. This trend holds for all the three DIF levels and for

Table 11. Power for the Long Common Block (20 Items), Impact Condition.

Item	DIF level	New	Current	Difference
1	Low	0.47	0.41	0.06
	Medium	0.89	0.86	0.03
	High	1.00	1.00	0.00
2	Low	0.26	0.23	0.03
	Medium	0.67	0.55	**0.11**
	High	0.98	0.94	0.03
3	Low	0.11	0.10	0.01
	Medium	0.24	0.17	0.06
	High	0.53	0.39	**0.14**
Average		0.57	0.52	0.05

Note: DIF = differential item functioning.

both the new and current methods. What is interesting is that looking at the column of difference, Item 3 shows the largest difference between the new and the current methods. In other words, the advantage of the new DIF detection method is more pronounced when the item is more difficult, a condition under which DIF detection methods in general struggle.

Table 6 shows that when the common block is short and when impact does exist, the new method leads to slightly higher Type I error than the current method. On average, the difference is about 2.0%. Also comparison against Table 4 shows that the Type I error is higher when impact is present, regardless of the difficulty level of the item and which method to use, which is expected.

Table 7 demonstrates that for the short common block and impact-present condition, the new method produces substantially *higher* power than the current method. On average, the difference is about 11%. Again as discussed previously, within each item, the higher the DIF level, the higher the power. Across items, the more difficult the item, the lower the power. Comparison of Table 5 against Table 7 shows that, as expected, the power of DIF detection methods is in general lower when impact exists. But the new method was more robust against the presence of impact. From Tables 5 through 7, by introducing impact, the new method only sees a reduction in power by 2%. In contrast, the current method takes a hard hit: The power is lower by 11%. As a result, the advantage of the new method over the current procedure is more pronounced when there is impact.

Table 8 shows the Type I error rates for the long common block (20 items) under the impact-free condition. Both DIF detection methods keep the Type I error well under control. The empirical Type I error rates are very close to the nominal level of 5%. The two methods are comparable in this regard. Comparison of Table 8 against Table 4 shows that longer common block results in lower Type I error. This holds regardless of item difficulty and the DIF detection method. This confirms the speculation that longer common block leads to better equating, and consequently more accurate matching, which in turn lowers Type I error.

Table 9 demonstrates that for the long common block and impact-free condition, the new method produces *higher* power than does the current method. On average, the difference is about 5%. Recall that when the common block is short and there is no impact, the advantage of the new method over the current method is 2%. When the common block is longer, the new method is more advantageous. Again, this might be attributed to the fact that longer common block leads to better equating.

Table 10 shows that, for the long common block, when there is impact, the new method leads to *comparable* Type I error rates than the current method does. Compared against Table 6, it is shown that long common block leads to lower Type I error.

Table 11 presents the power for the condition of long common block with impact. The new method is uniformly as powerful as or more powerful than the current method across all conditions. In some situations, the difference can be as large as 14%. On average, the difference is about 5%. The trends observed earlier regarding the relationship between power and DIF level and item difficulty still hold here. Compared against Table 9, the powers are uniformly lower. Clearly this is due to the presence of impact.

Summary and Discussion

Overall, the new equated pooled booklet method, while keeping the Type I error at a comparable level, leads to (sometimes substantially) higher power than the current method, regardless of the DIF level, item difficult level, and whether impact is present or not. Its advantage is particularly pronounced when the item is difficult, or when impact exists, or both. In all the power tables, conditions where the new method shows a margin of 10% or higher are marked with boldfaces. Under these conditions, the current method usually suffers, but the new method is able to maintain its power. So the new method comes in handy when the current approach struggles. Note that these power gains are obtained when the Type I error levels are generally comparable. Across all conditions, the difference in Type I error is no more than 2%. Therefore, it is a pronounced method for DIF detection for assessments that use multiple booklets that are anchored by a common set of items.

Several general trends are also observed. First of all, within each item, the higher the DIF level, the higher the power. Moreover, other things being equal, the more difficult the item, the lower the power. This holds for all the three DIF levels manipulated in this simulation. In addition, the Type I error rates under impact-free condition are lower than those under the impact-present condition, whereas the power in the former condition is higher. Last, the Type I error for the longer common block condition is lower than that for the long common block, and power is higher. This is attributable to more accurate equating when the booklets share more common items.

Note that the equating design can be used in conjunction with other DIF detection methods for tests with matrix sampled booklets as well. For example, current operational DIF detection methods also include M-H and standardization procedure. This study focuses on using the equated booklet method in conjunction with polySIBTEST. As the next step, the possibility of generalizing the equating/linking design to applications of M-H–based and standardization DIF detection procedures is explored. It is also worth noting that the proposed method here does not only apply to the BIB design, but it is also generally applicable to booklet design as long as the booklets share common blocks.

This study, however, is limited in several aspects. One limitation is the mechanism in which DIF is introduced. First, as mentioned earlier, both parallel DIF and directional DIF may be present in testing. The simulation study in this article concentrates on parallel DIF; that is, the difference in item characteristic curves between the reference and focal groups is introduced by a shift in the item difficulty parameter only. It will also be interesting to investigate the utility of equated booklet design in the detection of nonparallel DIF, which implies difference in the item discrimination parameters between two groups. Also, only Tucker's method is used here for equating; other equating methods such as the test characteristic curve (TCC) method by Stocking and Lord (1983), or concurrent calibration method, can be considered as well. As pointed out by a reviewer, position effect may exert itself because in various booklets, the same

block may appear in different positions (Frey et al., 2009). Taking the position effect into account, equating methods such as concurrent calibration maybe more appropriate. Because this article focuses on illustrating the utility of equating the booklets prior to DIF testing when multiple booklets are involved instead of comparing the effectiveness of different equating methods, only Tucker's method is implemented here. But it will be an interesting direction to follow-up. Another limitation is that only dichotomous items are simulated here. The polySIBTEST is completely capable of handing polytomous items. The proposed methodology is expected to work effectively with polytomous items as well. Finally, in most DIF analyses, the reference group tends to be much larger than the focal group. Future studies involving larger and unequal group sizes are warranted.

Acknowledgments

The authors would like to thank Dr. Swaminathan for his suggestions for this project and two anonymous reviewers for their constructive comments and feedback of the manuscript.

Declaration of Conflicting Interests

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

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by a grant from the U.S. Department of Education awarded to Hua-Hua Chang as principal investigator (PI) and University of Texas at Austin as grantee in 2002

Reference

Allen, N., Carlson, J., & Donoghue, J. (2001). Overview of Part II: The analysis of 1998 NAEP data. In N. Allen, J. Donoghue, & T. Schoeps (Eds.), *The NAEP 1998 technical report* (pp. 143-159). Washington, DC: National Center for Education Statistics.

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

Bolt, D., & Stout, W. (1996). Differential item functioning: Its multidimensional model and resulting SIBTEST detection procedure. *Behaviormetrika, 23,* 67-95.

Chang, H., & Mazzeo, J. (1994). The unique correspondence of the item response function and item category response functions in polytomous scored item response models. *Psychometrika*, *59*, 391-404.

Chang, 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.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the scholastic aptitude test. *Journal of Educational Measurement*, *23*, 355-368.

Dorans, N., Schmitt, A. P., & Bleistein, C. (1992). The standardization approach to assessing comprehensive differential item functioning. *Journal of Educational Measurement*, *29*, 309-319.

Frey, A., Hartig, J., & Rupp, A. (2009). An NCME instructional module on booklet designs in large scale assessments of student achievement: Theory and practice. *Educational Measurement: Issues and Practice*, *28*(3), 39-53.

Gonzalez, E., & Rutkowski, L. (2010). Principles of multiple matrix booklet designs and parameter recovery in large-scale assessments. In *IERI Monograph Series: Issues and Methodologies in Large-*

scale Assessments (*3rd Vol.*). Retrieved from http://www.ierinstitute.org/fileadmin/Documents/IERI_Monograph/IERI_Monograph_Volume_03_Chapter_6.pdf

Hanson, B. (1998). Uniform DIF and DIF defined by differences in item response functions. *Journal of Educational and Behavioral Statistics*, *23*, 244-253.

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. 129-146). Hillsdale, NJ: Lawrence Erlbaum.

Jenkins, F., & Kulick, E. (1994). Data analysis for the mathematics assessment. In E. Johnson & J. Carlson (Eds.), *The NAEP 1992 Technical Report* (pp. 299-341). Washington, DC: National Center for Education Statistics.

Johnson, E., & Carlson, J. (1994). *The NAEP 1992 Technical Report*. Washington, DC: National Center for Education Statistics.

Kolen, M., & Brennan, R. (1995). *Test equating, methods and practices*. New York, NY: Springer.

Lord, F. (1965). *Item sampling in test theory and in research design* (ETS Research Bulletin No. RB-65–22). Princeton, NJ: Educational Testing Service.

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.

Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics*, *17*(2), 131-154.

Qian, J., Steven, I., Lois, W., & Liang, J. (2001). Data analysis of the state reading assessment. In N. Allen, J. Donoghue, & T. Schoeps (Eds.), *The NAEP 1998 Technical Report* (pp. 307-344). Washington, DC: National Center for Education Statistics.

Robitaille, D. F., & Beaton, A. E. (2002). (Eds). *Secondary analysis of the TIMSS data*. Norwell, MA: Kluwer Academic.

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.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*, 201-210.

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

Swanson, D. B., Clauser, B. E., Case, S. M., Nungester, R. J., & Feathermean, C. (2002). Analysis of differential item functioning (DIF) using hierarchical logistic regression models. *Journal of Educational and Behavioral Statistics*, *27*, 53-75.

Yildirim, H. H., & Berberoglu, G. (2009). Judgmental and statistical DIF analyses of the PISA-2003 mathematics literacy items. *International Journal of Testing*, *9*, 108-121.

Zwick, R. (1987). Assessing the dimensionality of NAEP reading data. *Journal of Educational Measurement*, *24*, 293-308.