1-2016

The Matching Criterion Purification for Differential Item Functioning Analyses in a Large-Scale Assessment

Hyesun Lee
University of Nebraska-Lincoln, hlee7@unl.edu

Kurt F. Geisinger
University of Nebraska-Lincoln, kgeisinger2@unl.edu

Published in *Educational and Psychological Measurement* 76:1 (2016), pp. 141–163; doi: 10.1177/0013164415585166

The Matching Criterion Purification for Differential Item Functioning Analyses in a Large-Scale Assessment

HyeSun Lee and Kurt F. Geisinger

University of Nebraska-Lincoln

Corresponding author — Kurt F. Geisinger, University of Nebraska-Lincoln, 22G Teachers College Hall, Lincoln, NE 68588-0345, USA; email kgeisinger2@unl.edu

Abstract

The current study investigated the impact of matching criterion purification on the accuracy of differential item functioning (DIF) detection in large-scale assessments. The three matching approaches for DIF analyses (block-level matching, pooled booklet matching, and equated pooled booklet matching) were employed with the Mantel– Haenszel procedure. Five factors—the length of a test, the proportion of items exhibiting DIF, a sample size, a ratio of a reference and focal group, and the existence of an average ability difference between two groups—were manipulated. The three matching approaches were used with and without purification. Also, a systematic test form difference was considered. The results indicated that overall, matching criterion purification in the three approaches contributed to the improvement of power in the detection of DIF. Depending on the psychometric characteristics of items exhibiting DIF and the existence of an average ability difference, the amount of power improvement due to matching criterion purification was different across the three approaches. The purification of a matching criterion contributed to the slight reduction of Type I error rates in the three approaches when no mean ability difference existed between the two groups. Considering power improvement with the control of Type I error rates, the purification of a matching criterion in the pooled booklet matching and the equated pooled booklet matching approaches can be recommended for DIF analyses in large-scale assessments.

Keywords: differential item functioning (DIF), matching criterion purification, Mantel–Haenszel procedure, large-scale assessments

Overview

The consideration of test fairness is important in large-scale assessments that aim to compare educational achievement among various subgroups within a nation or across countries (Glas & Jehangir, 2014). A differential item functioning (DIF) analysis is one statistical approach to examine test fairness by identifying items that perform differentially across subgroups of test takers while controlling for test takers' ability.[1] Large-scale assessments need different approaches for DIF analyses due to the systematic sparseness of response data caused by multiple matrix sampling designs. Multiple matrix sampling designs, which are often employed in large-scale assessments such as the Programme for International Student Assessment (PISA) and the National Assessment of Educational Progress (NAEP), sample not only examinees from a population but also items from a total item pool (Rutkowski, Gonzalez, & von Davier, 2014). As illustrated in Table 1, the balanced incomplete block design that the PISA and the NAEP use is a specific type of multiple matrix sampling design. (For details about different types of matrix sampling designs, see Frey, Hartig, & Rupp, 2009.) The advantage of this multiple matrix sampling design is larger content domain coverage, while saving testing time by administering only a portion of item pools to each test taker (Goodman, Willes, Allen, & Klaric, 2011; Rutkowski et al., 2014); however, one of the disadvantages of this design is that traditional approaches for DIF analyses are not applicable due to the sparseness of responses.

DIF Analysis in a Large-Scale Assessment

Various statistical approaches for DIF analyses have been introduced over the past three decades: DIF methods based on item response theory (e.g., chi-square test [Lord, 1980; Thissen, Steinberg, & Wainer, 1993], Raju's area approaches [Raju, 1988, 1990]) and DIF methods based on classical test theory (e.g., the Delta plot approach [Angoff, 1982], the standardization approach [Dorans & Kulick, 1986], an approach based on logistic regression [Swaminathan & Rogers, 1990], and the Simultaneous Item Bias Test [Shealy & Stout, 1993]). Among these statistical methods, the Mantel–Haenszel (MH) procedure (Holland & Thayer, 1988) is one of the most widely used methods for DIF detection in most testing programs (French & Finch, 2013; Goodman et al., 2011) due to clear guidelines for reviewing and making decisions related to items exhibiting DIF (e.g., Zieky, 1993; Zwick, 2012; Zwick & Ercikan, 1989). Also, accumulated findings indicated the MH procedure performed better than other detection methods (e.g., Fidalgo, Ferreres, & Muniz, 2004; Hambleton & Rogers, 1989; Zwick, 1990). The MH procedure is the DIF detection method that many large-scale assessments including the NAEP and statewide achievement assessments employ (e.g., Educational Testing Service [ETS], 2006; National Center for Education Statistics [NCES], 2009).

The MH procedure flags an item as exhibiting DIF if the odds of getting an item correct significantly differ between a reference and focal group matched on their proficiency. To use the MH procedure for DIF analyses, a constant odds ratio (α_{MH}) across *n*-level of a matching criterion (e.g., total summed scores) is assumed. The chi-square test statistic is used to test the null hypothesis that the common odds ratio across all levels of a matching criterion equals to one with one degree of freedom. An estimate of the MH common odds ratio can be converted into a log odds ratio (Δ_{MH}, ETS delta scale), which is symmetric around zero. The delta scale shows a difference in item difficulty between two groups and can be employed to indicate the magnitude of DIF (see the Appendix for the computation of the α_{MH} and Δ_{MH} and the ETS categories of DIF magnitude).

To conduct DIF analyses, NAEP employs the MH procedure with *pooled booklet matching* as suggested by Allen and Donoghue (1996) (NCES, 2009). In the pooled booklet matching approach, total scores from pooled booklets are used as a matching criterion to control for test takers' proficiency. For instance, to conduct DIF analyses on items in Block A (see Table 1), total scores in each of three pooled booklets (Booklet 1, Booklet 2, and Booklet 3) that contain Block A are used as a matching criterion in the MH procedure. While a constant odds ratio across all levels of a matching criterion is assumed in the traditional MH procedure, the pooled booklet matching MH procedure assumes that the odds ratio is constant across all levels of the matching criterion scores across all pooled test booklets (Allen & Donoghue, 1996; Goodman et al., 2011). A common odds ratio across all levels of total scores in pooled booklets is tested for each item as done in the traditional MH procedure. The only difference between the traditional MH procedure and the pooled booklet matching MH procedure is that the latter employs total test scores across different tests (i.e., booklets) that contain the items for DIF analyses, while the former uses total test scores from only one test.

The advantage of the pooled booklet matching with the MH procedure is that this approach produces one DIF statistic for each item. Allen and Donoghue (1996) compared the pooled booklet matching with a booklet matching approach in which multiple DIF statistics for each item are produced. The *booklet matching* approach uses total scores from only one booklet as a matching criterion. Suppose that a DIF analysis for Item 1 in Block A (Table 1) is conducted based on the MH booklet matching approach. Since the booklet matching approach uses total scores from only one test booklet as a matching criterion, three separate DIF analyses need to be conducted by employing total scores from Booklet 1, Booklet 2,

Table 1. Balanced Incomplete Block Design

Booklet 1 (30 items)	Booklet 2 (30 items)	Booklet 3 (30 items)
Block A (10 items)	Block D (10 items)	Block F (10 items)
Block B (10 items)	**Block A (10 items)**	Block G (10 items)
Block C (10 items)	Block E (10 items)	**Block A (10 items)**

and Booklet 3, which contain Block A. Then, three DIF statistics for Item 1 in Block A, obtained from the MH DIF analyses with Booklet 1, Booklet 2, and Booklet 3, are combined for one DIF statistic of Item 1. According to Allen and Donoghue, because each of the three DIF statistics from the booklet matching approach is computed by using response data from only one booklet (relatively smaller sample size compared with the sample size from all of the three booklets), the DIF statistics tend to vary more than those estimated with a larger sample size from the three pooled booklets. That is, if the number of examinees taking each booklet was 100, then the three DIF statistics for Item 1 in Block A are estimated based on the sample size 100 from each of the three booklets; however, in the pooled booklet matching, the DIF statistic of Item 1 in Block A is estimated based on the sample size of 300, which can lead to more stable statistics than those based on the same size of 100.

To generate a single DIF statistic for each item another approach, named *block-level matching*, can be considered (Allen & Donoghue, 1996). In the MH block-level matching approach, DIF analyses are conducted by using total scores from each block as a matching criterion. For instance, total scores from Block A composed of 10 items (Table 1) are employed as a matching criterion in the MH block-level matching approach. Thus, the MH block-level matching approach allows one to conduct DIF analyses with larger sample sizes than the booklet matching approach. However, the use of total scores from only one block as a matching criterion can produce unreliable estimates of DIF statistics because the length of a block is much shorter (e.g., 10 items) than the length of a booklet (e.g., 30 items; Allen & Donoghue, 1996). With respect to a matching criterion from a relatively short test, Donoghue, Holland, and Thayer (1993) found that using total scores from fewer than 20 items as a matching criterion can threaten accuracy in the estimation of MH statistics. Zwick (1990) also stated that an unreliable matching criterion may jeopardize the MH procedure.

Comparing the three different matching approaches (the MH block-level matching, the MH booklet matching, and the MH pooled booklet matching), Allen and Donoghue (1996) found that the MH pooled booklet matching performed better in terms of controlling for Type I errors and detecting DIF (power improvement). Allen and Donoghue reported that the MH booklet matching showed a larger standard error in the DIF effect size (Δ_{MH}) than the MH pooled booklet matching.

Goodman et al. (2011) conducted an extended simulation study on the MH pooled booklet matching under three different booklet designs: balanced incomplete block design, common block design, and nonoverlapping matrix design. By comparing the Type I error rates, power levels, and the DIF statistics from the three booklet designs with those from a complete data set (no missing data), Goodman et al. found that the MH pooled booklet matching performed well in the three booklet designs when the sample size was large. That is, the sparseness did not affect the result of DIF detection when the MH pooled booklet matching was used under the condition of $N = 6{,}000$ and $N = 12{,}000$ sample sizes with equal and unequal ratios of focal and reference groups.

By emphasizing potential differences in the level of difficulty across test booklets, Cheng, Chen, Qian, and Chang (2013) suggested equated pooled booklet matching with the Simultaneous Item Bias Test procedure for both dichotomous and polytomous items (polySIBTEST; Chang, Mazzeo, & Roussos, 1996). The equated pooled booklet approach includes an additional equating step for DIF analyses after pooling test booklets for DIF analyses. The purpose of equating is to adjust potential differences in average difficulty across test booklets. The common items across pooled booklets (common block items, such as Block A in Table 1) are used as an anchor for equating, as in equating with the nonequivalent group design. Then, equated total scores, instead of raw total scores, are used as a matching criterion for the polySIBTEST. Cheng et al. compared the power levels and the Type I error rates of the equated pooled booklet matching approach to those of the booklet matching approach in which multiple DIF statistics from each booklet are produced. For the equating function, Tucker linear equating (Gulliksen, 1950) was used. Cheng et al. indicated that the equated pooled booklet matching approach showed slightly higher power than the booklet matching approach when the item identified as having DIF was more difficult and/or a mean ability difference between the reference and focal group existed. The Type I error rates were not significantly different between the two approaches. The aforementioned four different matching approaches, block-level matching, booklet matching, pooled booklet matching, and equated pooled booklet matching, are summarized in Table 2.

Purification of Matching Criterion

In addition to the decision for a matching approach in DIF analyses, large-assessment testing programs need to determine whether or not purification of a matching criterion will be employed. Matching criterion purification means the removal of items detected as DIF in a preliminary DIF analysis when computing matching criterion scores (total scores), thus allowing one to use only non-DIF items as

Table 2. Four Matching Approaches for Differential Item Functioning (DIF) Analyses in Large-Scale Assessments

Approach	Matching criterion
Block matching	Total scores from items in a block
Booklet matching	Total scores from items in a booklet
Pooled booklet matching	Total scores from items a pooled booklet that shares a block of items for DIF analyses with other booklets
Equated pooled booklet matching	Equated total scores based on total scores from items in a pooled booklet that share a block of items for DIF analyses; one booklet is used as a reference form and other booklets are new forms for equating

a matching criterion for the main DIF analyses (Clauser & Mazor, 1998; French & Maller, 2007; Holland & Thayer, 1988).

There are two approaches in conducting the purification of a matching criterion. One is the *two-step* procedure that Holland and Thayer (1988) suggested and the other is the *iterative* procedure. The difference between the two types of purification is the number of preliminary DIF analyses (DIF analyses conducted before a main DIF analysis) to filter out items flagged as DIF. If only one preliminary DIF analysis is conducted to remove DIF from the test, the procedure is called *two-step* purification (Holland & Thayer, 1988). If preliminary DIF analyses are conducted repeatedly until no items were flagged as DIF, this is the *iterative* purification (French & Maller, 2007).

Regarding the purification of a matching criterion in the MH procedure, Clauser, Mazor, and Hambleton (1993) found that the two-step purification improved overall power levels. When the proportion of DIF was large (20%) and the levels of mean ability between two groups were equal, purification of a matching criterion improved the power (ranging from 40% to 50%) to detect DIF. When there was a mean ability difference between two groups, purification of a matching criterion contributed to the improvement of power (22%) only with a relatively longer test (80 items). Matching criterion purification led to the reduction of Type I error rates in all test length conditions when the proportion of DIF was relatively large and the average levels of ability were equal; however, the reduction of Type I errors was observed only in the longer test with the larger proportion of DIF when a mean ability difference existed between the reference and focal group.

Also, Fidalgo, Mellenbergh, and Muniz (2000) examined the performance of the two purification types (two-step purification and iterative purification) including no purification in the MH procedure. Fidalgo et al. found that the iterative purification performed better than the other two purification approaches (two-step purification and no purification). In terms of power levels and Type I error rates, the iterative procedure performed well when the proportion of DIF was relatively large (15% and 30%). Additionally, Wang and Su (2004) found that the MH procedure with the twostep purification and the iterative purification performed better than no purification in most conditions; however, when the test length was short and the average level of test takers' ability differed between the focal and reference group, Type I error rates were increased regardless of the types of purification.

By employing the iterative purification, French and Maller (2007) examined the effects of purification in the logistic regression DIF method on power levels and Type I error rates. French and Maller stated that the iterative purification did not substantially contribute to the improvement of overall power and the control of Type I error rates in the logistic regression DIF method. Also, Magis and Facon (2012) examined whether the iterative purification in modified delta plot method affected power levels to detect DIF under small sample size conditions. Magis and Facon found that the iterative purification employed in the modified delta plot method did not contribute to the improvement of power levels to detect DIF.

Research Questions

Although many studies in the literature examined the effects of matching criterion purification on the detection of DIF, these studies were not focused on DIF analyses for large-scale assessments, but rather traditional DIF analyses in which all examinees take all items. In addition, the previous studies introducing different approaches for DIF analyses in large-scale assessments (e.g., Allen & Donoghue, 1996; Cheng et al., 2013) did not address issues related to the purification of a matching criterion. This void in the literature may make practitioners and applied researchers wonder whether purification of a matching criterion should be included for DIF analyses in large-scale assessments. Therefore, the current simulation study examined the effects of matching criterion purification on the detection of DIF in a large-scale assessment by employing the three different matching approaches: block-level matching, pooled booklet matching, and equated pooled booklet matching with the MH procedure. Of interest was whether the purification in the three matching approaches would improve power in the detection of DIF with the MH procedure, while controlling for Type I errors. The measures of accuracy in the detection of DIF were Type I error rates and power levels. The aim of the current study was to determine whether purification of a matching criterion is necessary for DIF analyses in large-scale assessments. Findings from the current study would be useful especially for testing programs in which scoring procedures should be quickly performed due to tight deadlines for reporting results (e.g., Miller & Fitzpatrick, 2009).

Method

Manipulated Factors

The current study manipulated five factors: the length of a test (30 items, 60 items, the lengths of the common blocks within tests were 10 items and 20 items, respectively), the proportion of items exhibiting DIF (10%, 20% of items in a common block), the sample size per booklet (400, 800), the ratio of a reference and focal group (1:1 and 3:1), and a difference in the average level of ability between a focal and reference group (equal, unequal). For equal mean ability conditions, $\sim N(0, 1)$ was used for both groups and for unequal mean ability conditions, $\sim N(0, 1)$ for the reference group, and $\sim N(-1, 0)$ for the focal group were used for the response data generation. Among the five factors, the length of a test, the proportion of items exhibiting DIF, and differences in the average level of ability between the two groups were based on the fact found these three factors were related to the effects of purification on DIF detection; Clauser et al. (1993) and Wang and Su (2004) found that the effects of purification were different depending on the length of a test and the existence of a mean ability difference between a reference and focal group; and Fidalgo et al.'s (2000) findings indicated that the proportion of items exhibiting DIF were related to the effects of matching criterion purification. The values of the manipulated factors were approximately emulated based on those employed in the

previous studies. The sample size per booklet was decided to mirror real large-scale assessment data: the average number of U.S. students per booklet participating in the PISA (NCES, n.d.). As a total 32 conditions were employed.

In addition to the five manipulated factors, differences in test booklet difficulty were considered in the current study. Differences in test booklet difficulty were implemented by differing difficulty parameters in a test booklet by 0.1. That is, except for the one booklet (Booklet 1) employed as a reference booklet, difficulty parameters for the other two booklets (Booklet 2 and Booklet 3) were adjusted either lower by 0.1 (Booklet 2, easy booklet) or higher by 0.1 (Booklet 3, difficult booklet). The adjusted difficulties were applied to only the nonanchor items (items in Blocks B, C, D, E, F, G in Table 1). The 0.1 difference in difficulty parameters led to approximately one tenth of a standard deviation difference in the level of test booklet difficulty. The one tenth of a standard deviation difference was chosen based on empirical test data examples in the literature (e.g., Kim, Livingston, & Lewis, 2011; Skaggs, 2005).

Procedure

Data Generation. Response data were generated by using the two-parameter logistic model instead of the three-parameter logistic model, based on the findings in the previous research; French and Finch (2013) stated that studies in the literature (e.g., Roussos & Stout, 1996) found that guessing parameters deleteriously affected the performance of the MH procedure when an average ability difference exists between a reference group and focal group. Three difficulty parameters (high, medium, and low) and three discrimination parameters (high, medium, and low) were used for items exhibiting DIF (Table 3). The parameters for items exhibiting DIF were taken from French and Finch (2013) and parameters for non-DIF items were from the 1998 NAEP reading for Grade 8 (Allen, Donoghue, & Schoeps, 2001). As indicated in Table 3, only uniform DIF, favoring a reference group, was considered. The magnitude of DIF was 0.4, which was computed based on the area between item characteristic curves. This magnitude corresponds to Category B in the ETS guideline (Zieky, 1993).[2] R (R Development Core Team, 2014) was used for data generation and analyses. The number of replications was 500.

Table 3. Parameters for Items Exhibiting DIF

	a-Parameter	b-Parameter (reference group)	b-Parameter (focal group)
DIF Item 1	1.25 (high a, med b)	–0.26	0.26
DIF Item 2	0.5 (low a, high b)	1.28	1.80
DIF Item 3	0.9 (med a, low b)	–1.80	–1.28
DIF Item 4	0.9 (med a, high b)	1.28	1.80

Analysis

Purification. The two-step procedure as in Holland and Thayer (1988) was employed in the current study. For the computation of purified scores, only true DIF items (items manipulated as DIF) were removed (a) to avoid the contamination of the purification effect due to Type I and/or Type II errors in DIF detection and (b) to keep the number of items in the common block constant because differing numbers of items in a common block would be a confounding factor to examine the effects of purification on DIF detection.

Data Analysis. The MH procedure with block-level matching, pooled booklet matching (Allen & Donoghue, 1996), and equated pooled booklet matching (Cheng et al., 2013) were employed with and without the purification of matching criterion scores for DIF analyses. Tucker linear equating was used for the equated booklet matching approach as was done in Cheng et al. (2013).

The impact of the matching criterion purification was examined in terms of power levels and Type I error rates. Power in the detection of DIF shows how accurately items manipulated as DIF were detected as DIF. Significance tests based on the MH chi-square test statistics (χ^2_{MH}) and the magnitude of DIF (Δ_{MH} corresponding to Category B) were used for the detection of DIF. Power for each item exhibiting DIF and the average level of power for all items exhibiting DIF were examined. Type I error rates indicate the proportion of times items without DIF were falsely detected as DIF. Mean Type I error rates across replications were reported. To examine whether purification influenced the improvement of power and the control of Type I error rates, marginal means of power and Type I error rates were also examined for each manipulated factor. Finally, ANOVAs were conducted to determine which manipulated factors affected the level of power and Type I error rates.

Results

Power

Overall, the pooled booklet matching and equated pooled booklet matching showed slightly higher power levels than the block-level matching approach in most conditions. With respect to the detection of DIF Item 1 (with high discrimination and medium difficulty, see Table 3), the power of all three approaches were more than 0.9 in all conditions. While purification of a matching criterion did not result in noticeable differences in terms of the power to detect the DIF Item 1, a distinguishable improvement (23%) in power was found with the use of purification when DIF Item 2 (with low discrimination and high difficulty) was included under the conditions where mean ability was different between a reference and focal group and the pooled booklet matching and equate pooled booklet matching approaches were employed. Interestingly, when the average ability level was equal, the block-level matching approach with purification performed as well as the other two approaches in detecting DIF Item 2.

When the length of a test was longer (60 items), the impact of purification on power levels in the pooled booklet matching and the equated pooled booklet matching appeared slightly smaller, by 3% on average, than the impact found in the shorter test (30 items). With the increase of the proportion of items exhibiting DIF, purification led to the increase of overall power by 17% in the pooled booklet matching and the equated pooled booklet matching, especially when the average ability was different between the two groups. The boldfaced numbers in Table 4 indicate more than 10% of power improvement due to the purification.

Sample size and ratio were not associated with the impact of purification on the power to detect DIF. Figure 1 shows these findings. Based on the results from ANOVA, the highest order significant interactions were the four-way interactions of a matching approach by purification by the proportion of DIF by mean ability ($F_{2,15} = 78.351$, $p = .000$, $\eta^2 = .913$) and a matching approach by purification by the length of a test by the proportion of DIF ($F_{2,15} = 42.815$, $p = .000$, $\eta^2 = .851$). These results indicated that the levels of power to detect DIF were significantly affected by the interactions of matching approaches and the purification and other two manipulated factors, the proportion of DIF and the existence of average ability difference.

As mentioned in the Method section, purified matching scores were computed by removing true DIF item(s) in the current study to eliminate confounding factors to investigate the impact of purification, which may not be always possible in real testing programs. By acknowledging that Type I and/or Type II errors in the detection of DIF may influence the purification process, additional analyses[3] were conducted to mirror real practice. That is, it was examined whether the purification would improve overall power, when (a) both true DIF items and falsely detected items were removed to purify matching criterion scores and (b) true DIF items were not removed for the computation of purified scores (Type II errors) and falsely detected items were removed to purify matching criterion scores (Type I errors). The results revealed that the inclusion of a falsely detected item, in addition to true DIF, did not change the amount of power improvement brought by the purification of matching criterion scores. Instead of a true DIF item, when only the falsely detected item was used for the purification of matching criterion scores, the overall power improvement was slightly lower than those reported in Table 4; however, the finding that the purification of a matching criterion contributed to the improvement of power still held.

Type I Error Rate

In general, Type I error rates with and without purification were under the nominal Type I error rate (less than 0.05) except for the block-level matching approach employed under the conditions where the average level of ability were unequal between the two groups. This finding that a mean ability difference between two groups was associated with the increase of Type I error rates was consistent with the findings in previous studies (e.g., Clauser et al., 1993). As expected, Type I error rates decreased as the sample size increased.

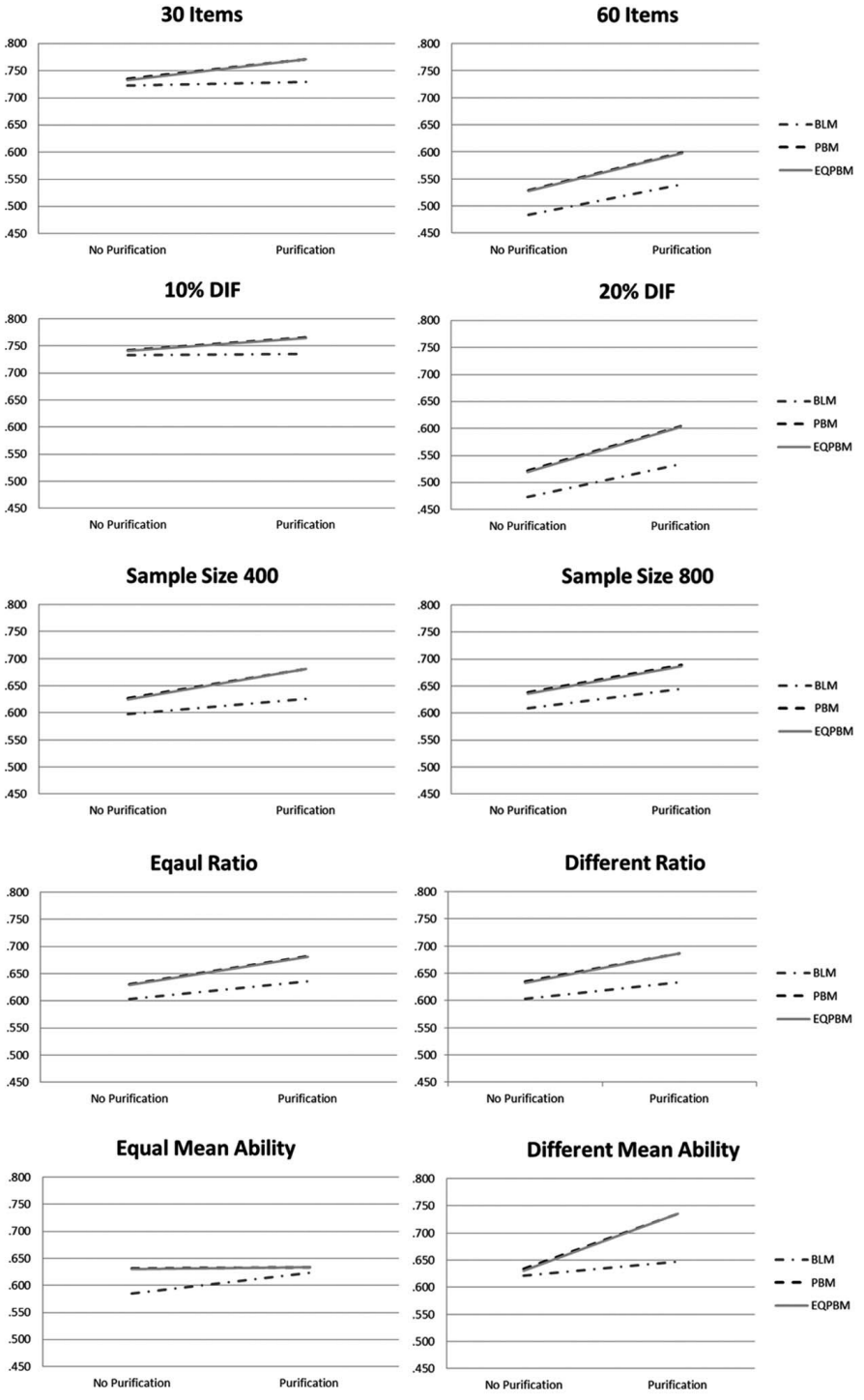

Figure 1. Impact of purification on power

Table 4. Power Levels.

	Sample size 400				Sample size 800			
	EAER	EADR	DAER	DADR	EAER	EADR	DAER	DADR
30 Items 10% DIF								
BLM	0.906	0.886	0.986	0.988	0.974	0.956	1	1
PBLM	0.892	0.886	0.990	0.980	0.966	0.954	1	0.998
PBM	0.944	0.916	0.960	0.946	0.980	0.982	0.992	0.986
PPBM	0.916	0.900	0.996	0.990	0.974	0.966	1	1
EQPBM	0.938	0.910	0.956	0.936	0.980	0.984	0.990	0.984
PEQPBM	0.918	0.898	0.996	0.992	0.976	0.964	**1**	1
30 Items 20% DIF								
BLM	0.455	0.479	0.489	0.492	0.489	0.467	0.498	0.500
PBLM	0.493	0.522	0.492	0.495	0.511	0.501	0.497	0.500
PBM	0.501	0.531	0.498	0.512	0.508	0.513	0.499	0.503
PPBM	0.504	0.532	**0.636**	**0.648**	0.509	0.512	**0.625**	**0.629**
EQPBM	0.501	0.534	0.498	0.506	0.508	0.513	0.499	0.500
PEQPBM	0.506	0.531	**0.633**	**0.655**	0.511	0.508	**0.619**	**0.627**
60 Items 10% DIF								
BLM	0.507	0.506	0.498	0.514	0.498	0.498	0.501	0.507
PBLM	0.519	0.525	0.506	0.522	0.508	0.508	0.504	0.515
PBM	0.519	0.537	0.514	0.537	0.511	0.514	0.507	0.537
PPBM	0.526	0.537	0.602	0.629	0.510	0.515	0.571	0.612
EQPBM	0.524	0.534	0.514	0.541	0.512	0.515	0.503	0.534
PEQPBM	0.525	0.532	0.597	0.632	0.512	0.517	0.567	0.609
60 Items 20% DIF								
BLM	0.428	0.439	0.487	0.496	0.443	0.427	0.495	0.494
PBLM	**0.540**	0.521	0.562	0.569	**0.587**	**0.545**	**0.617**	**0.601**
PBM	0.528	0.528	0.514	0.541	0.555	0.528	0.550	0.541
PPBM	0.549	0.552	**0.690**	**0.684**	0.582	0.554	**0.732**	**0.726**
EQPBM	0.526	0.528	0.522	0.532	0.554	0.525	0.541	0.537
PEQPBM	0.549	0.553	**0.691**	**0.690**	0.578	0.558	**0.727**	**0.722**

DIF = differential item functioning; EAER = equal ability equal ratio; EADR = equal ability different ratio; DAER = different ability equal ratio; DADR = different ability different ratio; BLM = block-level matching; PBLM = block-level matching with purification; PBM = pooled booklet matching; PPBM = pooled booklet matching with purification; EQPBM = equated pooled booklet matching; PEQPBM = equated pooled booklet matching with purification. The boldfaced values indicate the 10% or more power improvement due to the purification of a matching criterion.

With respect to the impact of purification on Type I error rates, the results demonstrated that the purification of a matching criterion reduced Type I error rates in the block-level matching under the condition of equal mean ability. In contrast, Type I error rates were not much different regardless of purification in the pooled booklet matching and the equated pooled booklet matching. That is, purification did not contribute to the reduction of the Type I error rates in all three DIF analysis approaches when the mean ability was different between the two groups (Table

Table 5. Type I Error Rates.

	Sample size 400				Sample size 800			
	EAER	EADR	DAER	DADR	EAER	EADR	DAER	DADR
30 Items 10% DIF								
BLM	0.0087	0.0238	0.0478	**0.0778**	0.0007	0.0013	0.0107	0.0258
PBLM	0.0071	0.0144	**0.0662**	**0.0804**	0.0000	0.0009	0.0367	0.0478
PBM	0.0038	0.0109	0.0167	0.0344	0.0002	0.0004	0.0009	0.0062
PPBM	0.0036	0.0118	0.0176	0.0347	0.0000	0.0004	0.0020	0.0067
EQPBM	0.0040	0.0109	0.0169	0.0349	0.0002	0.0004	0.0011	0.0062
PEQPBM	0.0036	0.0107	0.0164	0.0349	0.0000	0.0004	0.0016	0.0073
30 Items 20% DIF								
BLM	0.0218	0.0343	0.0493	**0.0835**	0.0008	0.0075	0.0098	0.0255
PBLM	0.0105	0.0188	**0.0513**	**0.0773**	0.0000	0.0013	0.0183	0.0290
PBM	0.0040	0.0133	0.0230	0.0433	0.0000	0.0010	0.0008	0.0073
PPBM	0.0025	0.0125	0.0205	0.0443	0.0000	0.0008	0.0030	0.0080
EQPBM	0.0040	0.0118	0.0223	0.0415	0.0000	0.0010	0.0005	0.0070
PEQPBM	0.0025	0.0118	0.0223	0.0450	0.0000	0.0008	0.0023	0.0078
60 Items 10% DIF								
BLM	0.0049	0.0110	0.0210	0.0393	0.0001	0.0008	0.0030	0.0077
PBLM	0.0029	0.0088	0.0304	0.0428	0.0001	0.0002	0.0118	0.0180
PBM	0.0027	0.0078	0.0123	0.0290	0.0000	0.0003	0.0010	0.0033
PPBM	0.0026	0.0069	0.0120	0.0300	0.0000	0.0003	0.0010	0.0042
EQPBM	0.0028	0.0080	0.0119	0.0280	0.0000	0.0003	0.0007	0.0032
PEQPBM	0.0027	0.0074	0.0127	0.0286	0.0000	0.0003	0.0008	0.0037
60 Items 20% DIF								
BLM	0.0096	0.0219	0.0210	0.0471	0.0006	0.0014	0.0040	0.0088
PBLM	0.0035	0.0111	0.0304	**0.0520**	0.0000	0.0000	0.0169	0.0264
PBM	0.0031	0.0099	0.0123	0.0274	0.0000	0.0001	0.0011	0.0030
PPBM	0.0021	0.0081	0.0120	0.0268	0.0000	0.0003	0.0013	0.0040
EQPBM	0.0028	0.0103	0.0119	0.0276	0.0000	0.0001	0.0013	0.0029
PEQPBM	0.0020	0.0091	0.0127	0.0281	0.0000	0.0003	0.0011	0.0040

DIF = differential item functioning; EAER = equal ability equal ratio; EADR = equal ability different ratio; DAER = different ability equal ratio; DADR = different ability different ratio; BLM = block-level matching; PBLM = block-level matching with purification; PBM = pooled booklet matching; PPBM = pooled booklet matching with purification; EQPBM = equated pooled booklet matching; PEQPBM = equated pooled booklet matching with purification. The boldfaced values indicate Type I error rates larger than 0.05.

5). This finding was consistent with those from previous research on purification with the MH procedure (e.g., Clauser et al., 1993; Wang & Su, 2004) and with the logistic regression approach (e.g., French & Maller, 2007). The highest Type I error rates of all three matching approaches were found under the conditions with the small sample size (400), the existence of mean ability difference between the two groups, and an unequal ratio of the two groups. The boldfaced values in Table 5 indicate Type I error rates larger than 0.05.

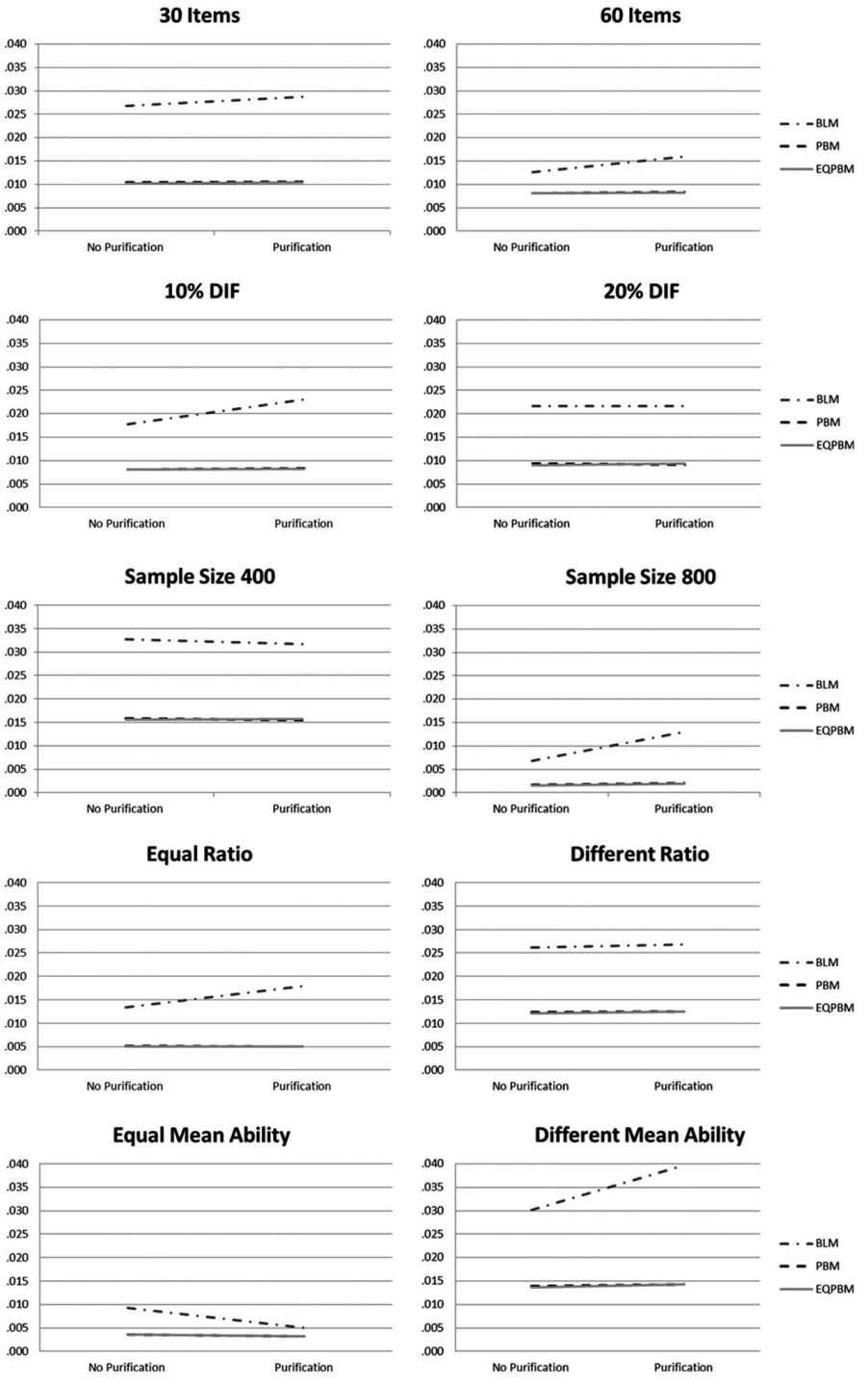

Figure 2. Impact of purification on Type I errors

When the test length was increased from 30 to 60 items, the overall Type I error rates were decreased, as found in Clauser et al. (1993). When it comes to the impact of purification on Type I error rates, however, the length of a test was not associated with the impact of purification on Type I error rates in the pooled booklet matching and the equated pooled booklet matching. Also, the proportion of DIF, ratio of the two groups, sample size did not influence Type I error rates in the pooled booklet matching and equated pooled booklet matching approaches. In the block-level matching approach, the purification of a matching criterion actually led to the increase of Type I error rates in many conditions except for equal ability conditions. Figure 2 displays these findings. In terms of the reduction of Type I error rates, purification was beneficial only with the use of block-level matching under the condition where the mean ability was equal between the reference and focal groups. Based on the results from ANOVA, the highest order significant interactions were the five-way interactions of a matching approach by purification by the proportion of DIF by the length of a test by sample size ($F_{2,15} = 4.015$, $p = .040$, $\eta^2 = .349$) and a matching approach by purification by the length of a test by the proportion of DIF by mean ability ($F_{2,15} = 3.846$, $p = .045$, $\eta^2 = .339$). These results indicated that the levels of Type I error rates were significantly affected by the interactions of matching approaches and the purification and other four manipulated factors, the proportion of DIF, the length of a test, the size of a sample, and the existence of average ability difference.

Type I error rates in the current study were relatively small; therefore, additional analyses were conducted with the conditions where no items exhibit DIF (null conditions). By employing the null conditions, whether or not the low rates of Type I errors were related to the existence of DIF could be examined. The Type I error rates from the null conditions were small (less than 0.05) and showed quite similar patterns as in Table 5. Also, the differences in Type I error rates between the null conditions and DIF conditions were minute. An interesting finding was that relatively high Type I error rates were detected for the item with high discrimination and high difficulty, especially under the null conditions where the average ability was different between the reference and focal group, the ratio of the two groups was unequal, and the sample size was small.

Discussion

Results from large-scale assessments have consequential impacts on teaching, learning, evaluation of educational systems, and policies in education (Gilmore, 2002; van den Heuvel-Panhuizen, Robitzsch, Treffers, & Köller, 2009). Thus, validity of test scores should be ensured and the comparability of scores across subgroups of test takers needs to be thoroughly examined (Glas & Jehangir, 2014). Conducting traditional DIF analyses as a part of validation procedures, however, is quite challenging due to the systematically missing data embedded in the design of large-scale assessments (Allen & Donoghue, 1996; Goodman et al., 2011). Also, considering

that many testing programs should quickly perform the scoring of tests due to tight deadlines for reporting results (Miller & Fitzpatrick, 2009) and the MH procedure is one of the DIF methods that large-scale assessments often employ (e.g., NAEP), informing test practitioners whether purification in the MH procedure for large-scale assessments improves the accuracy of DIF detection can be useful.

Regarding the research question—*does the matching criterion purification in the three different DIF analyses approach with the MH procedure contribute to the increase of the power while controlling for the inflation of the Type I error rate?*— the current simulation study found that in general purification contributed to the improvement of power in the detection of DIF. The overall power improvement due to purification in all three approaches was 5.3%; the average power improvement was 5.4% in the block-level matching, 5.2% in the pooled booklet matching, and 5.4% in the equated pooled booklet matching across all conditions. Depending on the psychometric characteristics of items exhibiting DIF and the existence of an average ability difference between the reference and focal groups, power improvement due to purification was different across the three DIF analysis approaches.

When the difficulty of the item flagged as DIF was medium and the discrimination was high (DIF Item 1), power improvement due to purification was quite small across all three DIF analysis approaches. However, under the conditions in which the difficulty of the item with DIF was an extreme value (either high or low, DIF Items 2, 3, 4), the matching criterion purification contributed to the most improvement of power, especially when the pooled booklet matching and the equated pooled booklet matching were employed with the existence of an ability difference.

Focusing on the impact of purification on the detection of DIF Item 2 (DIF with low discrimination and high difficulty parameters), purification improved power more under the condition where a relatively larger proportion of items exhibit DIF. This result was consistent with the findings that power improvement due to purification was larger when the proportion of DIF was larger (Clauser et al., 1993; French & Maller, 2007). Also, power improvement in the detection of DIF Item 2 (due to the purification with the pooled booklet matching and the equated pooled booklet matching) was larger under the condition where the length of a test was relatively shorter. Finally, the results from the current study also indicated that the effect of purification on power did not seem to be affected by the ratio of the reference and the focal groups.

With respect to Type I error rates, purification contributed only to a small amount of reduction in Type I error rates (only by 0.1%) when no ability difference existed. The most reduction with the use of purification (by 0.3% on average across all conditions) was observed in the block-level matching, while the purification of a matching criterion did not reduce Type I error rates in the other two approaches. When there was a mean ability difference between the two groups, purification slightly increased Type I error rates across all three DIF analysis approaches on average by 0.3%. However, the amount of increase in Type I error rates in the pooled booklet matching or the equated pooled booklet matching due to purification was very small, whereas the greatest increase

was found in the block-level matching; on average 0.9% increase was detected across all conditions.

Type I error rates were all under the nominal level of 0.05 in the pooled booklet matching and the equated pooled booklet matching, regardless of the purification process. These low values may raise a question as to why these two approaches were conservative when flagging items as DIF. Using summed scores based on more items (30 or 60 items in a pooled booklet) than are used for the DIF analyses (10 or 20 items in a common block) may mirror the effect of employing longer tests. That is, as found in the previous research (e.g., Clauser et al., 1993), Type I error rates in longer tests were smaller than in shorter test. Accordingly, it can be hypothesized that these two approaches' matching criterion scores, computed based on the relatively longer test, were more stable, which may lead to lower Type I error rates compared with the traditional MH procedure in which matching criterion scores are not computed from a longer test. However, this hypothesis was not examined in the current study. Additional research is necessary to investigate whether or not this hypothesis holds.

In sum, purification inflated the Type I error rate, only when the block-level matching was employed under the ability difference condition with relatively smaller sample size (400 sample size condition), otherwise, purification either decreased the Type I error rate by a small amount or showed similar Type I error rates as those with no purification.

The power improvement should be considered together with the inflation of Type I error rates (Cheng et al., 2013; Clauser et al., 1993). Based on the findings from the current research, purification of a matching criterion for DIF analyses in large-scale assessments can be recommended when a mean ability difference is more likely to exist between reference and focal groups and the pooled booklet matching or the equated pooled booklet matching is used for DIF analyses. Even though the purification in the pooled booklet matching or in the equated pooled booklet matching may slightly increase Type I error rates when an average ability difference existed between two groups, the Type I error rates were still less than 0.05. Therefore, the employment of purification for DIF analyses would be beneficial, resulting in the improvement of power for the detection of DIF. Considering that the difference in ability between the reference and focal groups is more common in practice (Clauser et al., 1993; French & Maller, 2007; Narayanan & Swaminathan, 1994), the use of purification with either the pooled booklet matching or the equated pooled booklet matching approaches in large-scale assessments would bring a practical benefit to test practitioners, since the two-step purification with the MH procedure can be implemented easily with various statistical programs without much effort. When the ability levels of the reference and focal groups were equal, the use of purification in the pooled booklet matching and the equated pooled booklet matching did not seem to improve power; however, lower Type I error rates in both approaches were detected with the use of purification. Thus, the use of purification in the two approaches may still be beneficial for the DIF analyses with subgroups of equal mean ability.

Additionally, the pooled booklet matching and the equated pooled booklet matching performed better than the block-level matching among the three matching approaches with the MH procedure. To detect DIF in items with the extreme difficulty parameters (high or low) when ability differences also existed, the pooled booklet matching and the equated pooled booklet matching outperformed the blocklevel matching. Related to the performance of the equated pooled booklet matching, the current study supported what Cheng et al. (2013) found in their simulation study; the equated pooled booklet matching performed well in the detection of items with DIF when item difficulty was high. Also, relatively higher Type I error rates in the block-level matching (than other two approaches) that the current study found was in line with Allen and Donoghue (1996).

One interesting finding in the present study was that the pooled booklet matching and the equated pooled booklet matching (regardless of the purification) showed similar performance in terms of power and the Type I error rate. At the time of writing this article, no other study has compared pooled booklet matching and equated pooled booklet matching. Related to the equated pooled booklet method, Cheng et al. (2013) reported that the equated pooled booklet matching with the polySIBTEST performed better than the booklet matching where multiple DIF statistics from separate DIF analyses are combined into one. Cheng et al. differed from the present study in several ways: (a) different DIF analysis approaches, (b) different sample sizes per booklet, (c) no consideration on the unequal ratio between the reference and focal group, and (d) no consideration on systematic form differences across booklets. Thus, the findings from Cheng et al. and that from the current study cannot be directly compared. However, considering that the current study considered systematic test form differences often found in large-scale assessments, it would be a practical implication that the Tucker linear equating in the pooled booklet matching might not be necessary when test form differences are within one tenth of a standard deviation, which will allow testing programs to save time scoring tests. This is because the MH procedure needs whole-numbered scale scores for the matching criterion, and equated scores from the Tucker linear equating function must be rounded to become whole numbers. As a result, the benefit from the linear equating to adjust the test form difference may be cancelled out due to the rounding. However, since the current study used the MH procedure only with Tucker linear equating for the equated pooled booklet matching, other DIF analysis methods with different equating functions need to be examined for the generalization of the current finding.

Finally, there are some limitations in the present research. The purified scores were computed by removing true DIF. That is, the removed items to purify the matching criterion were predetermined. Since this was a simulation study, it was possible to remove items with true DIF from the test when the purification was conducted; however, this is rarely possible in practice; thus, the items employed for the purification of a matching criterion may possibly be flagged due to Type I errors and should have not been excluded. As previously mentioned in the Results section, additional analyses revealed that items falsely flagged as DIF did not

affect the effect of purification, whereas Type II errors affected the improvement of power brought by purification. Thus, future research on the association of Type II errors and the impact of purification would be required for the generalization of the current findings. Also, only one equating function was employed for the equated pooled booklet matching. Further examination needs to be conducted by using different equating functions under various systematic test form difference conditions. Finally, the position effect of the common block was not considered in the current study. Therefore, future research should consider these limitations to enhance the generalizability of the current findings.

Appendix

The Mantel–Haenszel Procedure for Differential Item Functioning Analyses

- Null hypothesis

$$H_0 = \frac{\mathrm{FR}_{\mathrm{ref}(n)}}{\mathrm{FW}_{\mathrm{ref}(n)}} = \alpha\left[\frac{\mathrm{FR}_{\mathrm{foc}(n)}}{\mathrm{FW}_{\mathrm{foc}(n)}}\right], \quad \alpha = 1, n = 1, \ldots, N$$

$$H_1 = \frac{\mathrm{FR}_{\mathrm{ref}(n)}}{\mathrm{FW}_{\mathrm{ref}(n)}} = \alpha\left[\frac{\mathrm{FR}_{\mathrm{foc}(n)}}{\mathrm{FW}_{\mathrm{foc}(n)}}\right], \quad \alpha \neq 1,\ n = 1, \ldots, N \tag{1}$$

(α = common odds ratio, N = all levels of a matching criterion)

- Test the Null hypothesis (H_0: $\alpha = 1$)

$$\mathrm{MH} - \chi^2 = \frac{\left[\left|\sum_1^n \mathrm{FR}_{\mathrm{ref}(n)} - \sum_1^n \mathrm{E}(\mathrm{FR}_{\mathrm{ref}(n)})\right| - 0.5\right]^2}{\sum_1^n \mathrm{Var}(\mathrm{FR}_{\mathrm{ref}(n)})} \tag{2}$$

$$\mathrm{E}\left(\mathrm{FR}_{\mathrm{ref}(n)}\right) = \mathrm{E}\left(\mathrm{FR}_{\mathrm{ref}(n)} | \alpha_n = 1\right)$$

$$= \frac{\left[\left(\mathrm{FR}_{\mathrm{ref}(n)} + \mathrm{FW}_{\mathrm{ref}(n)}\right)^* \left(\mathrm{FR}_{\mathrm{ref}(n)} + \mathrm{FR}_{\mathrm{foc}(n)}\right)\right]}{\left[\mathrm{FR}_{\mathrm{ref}(n)} + \mathrm{FW}_{\mathrm{ref}(n)} + \mathrm{FR}_{\mathrm{foc}(n)} + \mathrm{FW}_{\mathrm{foc}(n)}\right]} \tag{3}$$

$$\mathrm{Var}\left(\mathrm{FR}_{\mathrm{ref}(n)}\right) = \mathrm{Var}\left(\mathrm{FR}_{\mathrm{ref}(n)} | \alpha_n = 1\right)$$

$$= \frac{\left(\mathrm{FR}_{\mathrm{ref}(n)} + \mathrm{FW}_{\mathrm{ref}(n)}\right)^* \left(\mathrm{FR}_{\mathrm{ref}(n)} + \mathrm{FR}_{\mathrm{foc}(n)}\right)^* \left(\mathrm{FR}_{\mathrm{foc}(n)} + \mathrm{FW}_{\mathrm{foc}(n)}\right)^* \left(\mathrm{FW}_{\mathrm{ref}(n)} + \mathrm{FW}_{\mathrm{foc}(n)}\right)}{\left(\mathrm{FR}_{\mathrm{ref}(n)} + \mathrm{FW}_{\mathrm{ref}(n)} + \mathrm{FR}_{\mathrm{foc}(n)} + \mathrm{FW}_{\mathrm{foc}(n)}\right)^{2} * \left(\left(\mathrm{FR}_{\mathrm{ref}(n)} + \mathrm{FW}_{\mathrm{ref}(n)} + \mathrm{FR}_{\mathrm{foc}(n)} + \mathrm{FW}_{\mathrm{foc}(n)}\right) - 1\right)} \tag{4}$$

- Estimate of a common odds ratio

$$\alpha_{MH} = \frac{\sum_1^n \left\{ \frac{(FR_{ref(n)} * FW_{foc(n)})}{(FR_{ref(n)} + FW_{ref(n)} + FR_{foc(n)} + FW_{foc(n)})} \right\}}{\sum_1^n \left\{ \frac{(FR_{foc(n)} * FW_{ref(n)})}{(FR_{ref(n)} + FW_{ref(n)} + FR_{foc(n)} + FW_{foc(n)})} \right\}} \quad (5)$$

$$\Delta_{MH} = -2.35 \ln(\alpha_{MH}) \quad (6)$$

$FR_{ref(n)}$ = Frequency of getting an item right at each score in a reference group
$FW_{ref(n)}$ = Frequency of getting an item wrong at each score in a reference group
$FR_{foc(n)}$ = Frequency of getting an item right at each score in a focal group
$FW_{foc(n)}$ = Frequency of getting an item wrong at each score in a focal group

Declarations — The authors received no financial support for the research, authorship, or publication of this article and declared no potential conflicts of interest with respect to its research, authorship, or publication.

Notes

1. DIF is a necessary but not the sufficient indication of bias (Clauser & Mazor, 1998; Goodman, Willes, Allen, & Klaric, 2011; Longford, 2014).
2. According to the ETS guidelines (Zieky, 1993), items are assigned one of three categories based on the Δ_{MH} (MH-Delta, difference in items difficulty between the two groups). Based on the absolute values of Δ_{MH} and chi-square significance test results, items are classified into Category A (negligible DIF, when Δ_{MH} is not significantly different from zero or the absolute value Δ_{MH} of is less than 1.0), Category B (slight to moderate DIF, when $\Delta_{\alpha MH}$ is significantly different from zero and the absolute value of Δ_{MH} is at least 1.0 and does not meet the criterion of Category C), or Category C (moderate to large DIF, when $\Delta_{\alpha MH}$ is significantly greater than 1.0 and the absolute value of Δ_{MH} is 1.5 or more). Items identified as Category C (or Category B that is close enough to Type C) are usually considered for item review (Zieky, 1993).
3. Two scenarios in which (a) only a falsely detected item was removed for the purification of a matching criterion and (b) both a falsely detected item and a true DIF item were removed for matching criterion purification were examined under the simulation condition where the largest power improvement due to purification was found; the length of a test is short (30 items), sample size is small (400 examinees), a mean ability difference exists between a reference and focal group, and the ratio of the two groups is unequal.

References

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

Allen, N. L., Donoghue, J. R., & Schoeps, T. L. (2001). *The NAEP 1998 technical report* (NCES 2001-509). Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics.

Angoff, W. H. (1982). Use of difficulty and discrimination indices for detecting items bias. In R. A. Berk (ed.), *Handbook of methods for detecting test bias* (pp. 96-116). Baltimore, MD: Johns Hopkins University Press.

Chang, H., Mazzeo, J., & Roussos, L. (1996). Detecting DIF for polytomous scored items: An adaptation of the SIBTEST procedure. *Journal of Educational Measurement*, *33*, 333-353.

Cheng, Y., Chen, P., Qian, J., & Chang, H. (2013). Equated pooled booklet method in DIF testing. *Applied Psychological Measurement*, *37*, 276-288.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differential functioning test items. *Educational Measurement: Issues and Practice*, *17*(1), 31-44.

Clauser, B. E., Mazor, K. M., & Hambleton, R. K. (1993). The effects of purification of the matching criterion on the identification of DIF using the Mantel-Haenszel procedure. *Applied Measurement in Education*, *6*, 269-279.

Donoghue, J. R., Holland, P. W., & Thayer, D. T. (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. 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.

Educational Testing Service. (2006). *California Alternate Performance Assessment (CAPA) technical report*. Online http://www.cde.ca.gov/ta/tg/sr/documents/capatechreport.pdf

Fidalgo, A., Ferreres, D., & Muniz, J. (2004). Liberal and conservative differential item functioning detection using Mantel-Haenszel and SIBTEST: Implications for Type I and Type II errors rates. *Journal of Experimental Education*, *73*, 23-29.

Fidalgo, A. M., Mellenbergh, G. J., & Muniz, J. (2000). Effects of amount of DIF, test length, and purification type on robustness and power of Mantel-Haenszel procedures. *Methods of Psychological Research Online*, *5*, 43-53.

French, B. F., & Finch, W. H. (2013). Extension of Mantel-Haenszel for multilevel-DIF detection. *Educational and Psychological Measurement*, *73*, 648-671.

French, B. F., & Maller, S. J. (2007). Iterative purification and effect size use with logistic regression for differential item functioning detection. *Educational and Psychological Measurement*, *67*, 373-393.

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

Gilmore, A. (2002). Large-scale assessment and teachers' assessment capacity: Learning opportunities for teachers in the National Education Monitoring Project in New Zealand. *Assessment in Education*, *9*, 343-361.

Glas, C., & Jehangir, K. (2014). Modeling country-specific differential item functioning. In L. Rutkwoski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment* (pp. 97-115). Boca Raton, FL: CRC Press.

Goodman, J. T., Willes, 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.

Gulliksen, H. (1950). *Theory of mental tests*. New York, NY: Wiley.

Hambleton, R., & Rogers, H. J. (1989). Detecting potential biased test items: Comparison of IRT and Mantel-Haenszel methods. *Applied Measurement in Education*, *2*, 313-334.

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

Kim, S., Livingston, S. A., & Lewis, C. (2011). Collateral information for equating small samples: A preliminary investigation. *Applied Measurement in Education*, *24*, 302-323.

Longford, N. T. (2014). Screening test items for differential item function. *Journal of Educational and Behavioral Statistics*, *39*, 3-21.

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

Magis, D., & Facon, B. (2012). Item purification does not always improve DIF detection: A counterexample with Angoff's delta plot. *Educational and Psychological Measurement*, *73*, 293-311.

Miller, G. E., & Fitzpatrick, S. J. (2009). Expected equating error resulting from incorrect handling of item parameter drift among the common items. *Educational and Psychological Measurement*, *69*, 357-368.

Narayanan, P., & Swaminathan, H. (1994). Performance of the Mantel-Haenszel and simultaneous item bias procedures for detecting differential item functioning. *Applied Psychological Measurement*, *18*, 315-328.

National Center for Education Statistics. (2009, March 18). *NAEP technical documentation: The Mantel-Haenszel procedure*. Online http://nces.ed.gov/nationsreportcard/tdw/analysis/scaling_checks_dif_proced_mh.aspx

National Center for Education Statistics. (n.d.). *Program for International Student Assessment (PISA): Frequently asked questions*. Online http://nces.ed.gov/surveys/pisa/faq.asp

R Development Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Online http://www.rproject.org

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, *53*, 495-502.

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement*, *14*, 197-207.

Roussos, L. A., & Stout, W. F. (1996). Simulation studies of the effect of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel Type I error performance. *Journal of Educational Measurement*, *33*, 215-230.

Rutkowski, L., Gonzalez, E., & von Davier, M. (2014). Assessment design for international large-scale assessments. In L. Rutkwoski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment* (pp. 75-95). Boca Raton, FL: CRC Press.

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

Skaggs, G. (2005). Accuracy of random groups equating with very small samples. *Journal of Educational Measurement*, *42*, 309-330.

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. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (eds.), *Differential item functioning* (pp. 67-113). Hillsdale, NJ: Erlbaum.

van den Heuvel-Panhuizen, M., Robitzsch, A., Treffers, A., & Köller, O. (2009). Large-scale assessment of change in student achievement: Dutch primary school students' results on written division in 1997 and 2004 as an example. *Psychometrika*, *74*, 351-365.

Wang, W.-C., & Su, Y.-H. (2004). Effects of average signed area between two item characteristic curves and test purification procedures on the DIF detection via the Mantel-Haenszel method. *Applied Measurement in Education*, *17*, 113-144.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland & H. Wainer (eds.), *Differential item functioning* (pp. 337-347). Hillsdale, NJ: Erlbaum.

Zwick, R. (1990). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide? *Journal of Educational Statistics*, *15*, 185-197.

Zwick, R. (2012). *A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement* (ETS RR-12-08). Princeton, NJ: ETS. Online http://www.ets.org/research/contact.html

Zwick, R., & Ercikan, K. (1989). Analysis of differential item functioning in the NAEP history assessment. *Journal of Educational Measurement*, *26*, 55-66.