Research Report

The Value of the Studied Item in the Matching Criterion in Differential Item Functioning (DIF) Analysis

Xuan Tan

Bihua Xiang

Neil J. Dorans

Yanxuan Qu

April 2010

ETS RR-10-13

The Value of the Studied Item in the Matching Criterion in Differential Item Functioning (DIF) Analysis

Xuan Tan, Bihua Xiang, Neil J. Dorans, and Yanxuan Qu
ETS, Princeton, New Jersey

April 2010

Technical Review Editor: Dan Eignor

Technical Reviewers: Longjuan Liang and Jinghua Liu

Abstract

The nature of the matching criterion (usually the total score) in the study of differential item functioning (DIF) has been shown to impact the accuracy of different DIF detection procedures. One of the topics related to the nature of the matching criterion is whether the studied item should be included. Although many studies exist that suggest the studied item should always be included in the criterion, the validity of this statement for models other than the Rasch model has not been studied. This study evaluates the effect of including/excluding the studied item in the matching criterion for situations that mimic real testing situations where the assumptions of the Rasch model are violated. A simulation study was conducted where the effect of including/excluding the studied item in the matching criterion was studied relative to different magnitudes of DIF and different group ability distributions, for data that follow the two-parameter logistic (2PL) item response theory (IRT) and multidimensional item response theory (MIRT) models. Results from the study show that including the studied item leads to less biased DIF estimates and more appropriate Type I error rate, especially when the group ability distributions are different. Systematic bias positioning DIF estimation in favor of the high ability group was consistently found across all simulated conditions when the studied item was excluded from the matching criterion.

Key words: differential item functioning, item inclusion, matching criterion, Mantel-Haenszel statistic, standardization

Table of Contents

Page

List of Tables

Theoretical Framework and Objectives

Analysis of differential item functioning (DIF), which provides an indicator of the presence of construct-irrelevant variance, is an important aspect in ensuring fairness and equity in testing. DIF occurs when different groups of examinees with the same level of proficiency on a test have different probabilities of answering an item correctly. There is a vast amount of research literature on this topic. Studies have been conducted to assess DIF across gender, ethnicity, language, and other demographic groups (e.g., Hamilton, 1999). A number of statistical procedures used commonly today include the standardization procedure (Dorans & Kulick, 1986), the simultaneous item bias test (SIBTEST) procedure (Shealy & Stout, 1993), the Mantel-Haenszel statistic (Holland & Thayer, 1988), item response theory (IRT)-based procedures (Thissen, Steinberg, & Wainer, 1993), and logistic regression (Swaminathan & Rogers, 1990). Different analysis procedures were also proposed to reduce bias and estimation error through iterative purification procedures to remove DIF items from the matching criterion (e.g., Zenisky, Hambleton, & Robin, 2003). The nature of the matching criterion that is used as a surrogate for examinee ability has been shown to impact the accuracy of the different DIF procedures in identifying DIF items. One issue related to this is whether the studied item should be included in the matching criterion.

This topic was investigated by many researchers in the late 1980s to early 1990s. Holland and Thayer (1988) illustrated mathematically that when the data were consistent with the Rasch model (Lord & Novick, 1968), inclusion of the studied item in a purified rights-scored matching criterion was necessary to avoid biased estimates of DIF for the studied item. Inclusion of the studied item removes the dependence of the item response on group differences in ability distributions. This means that, when there is no DIF present in the studied item, inclusion of the item in the matching criterion will always give a Mantel-Haenszel (MH) odds ratio of 1, while exclusion of the item will give an MH odds ratio other than 1, and will indicate DIF that is favoring the high ability group (positive DIF when the focal group has higher ability and negative DIF when the reference group has higher ability).

Zwick (1990) and Lewis (1993) developed this idea further to illustrate the applicability of this finding to more general item response models. Both authors proved mathematically that the bias correction benefit of including the studied item in the matching criterion held true for the binomial model, and claimed that for any IRT model more complex than the Rasch model, the advantage of including the studied item in the matching criterion would not be evident. Two other studies conducted simulation work and concluded that the studied item should always be included in the matching criterion to reduce bias in DIF estimates, even for more complex models. Donoghue, Holland, and Thayer (1993) evaluated the effect of including/excluding the studied item under the three-parameter logistic (3PL) model as a part of their study. Data simulation in their study was done by setting the discrimination parameters (*a*) to be fixed for all items in a simulated test in each studied condition, setting the guessing (*c*) parameter fixed for all conditions, and varying the difficulty parameters for different items for each studied condition. Although the 3PL model (Lord & Novick, 1968) was used to simulate data, only the *b* parameters were allowed to vary, which is a highly restricted form of the 3PL model.

Zwick, Donoghue, and Grima (1993) studied this issue by extending the scope of their DIF study to performance tasks. In their study, multiple-choice (MC) items and performance tasks were simulated using the 3PL model and the partial credit model respectively. The MC items were simulated to be free of DIF and were used as the matching criterion. The performance tasks were simulated to be the studied items with or without DIF. Although the MC items were simulated using the 3PL model, the studied items were always simulated with the partial credit model (Masters, 1982) that is an extension of the Rasch model for dichotomous items. Since the focus was on inclusion/exclusion of the studied item, this study was basically extending use of the Rasch model for dichotomous items to use of the model with polytomous items. Thus, the question of whether the advantage of including the studied item in the matching criterion would still hold for more complex models remains unanswered.

Zwick (1990) illustrated theoretically that for models more complex than Rasch, including the studied item would lead to random bias, indicating some null DIF items that favor the high ability group and some that favor the low ability group, but excluding the studied item would always lead to bias favoring the high ability group. How well this result can be replicated in practical applications is unknown. The advantage of including the studied items for conditions that depart from the Rasch model still needs to be investigated empirically. For example, more realistically, a test could have items with different discrimination parameters. Another interesting and applicable departure from the Rasch model is multidimensionality. When a test is multidimensional, the inclusion/exclusion of the studied item might have differential impact for items belonging to different dimensions.

The purpose of this study was to study the impact of including/excluding the studied item in the matching variable on bias in DIF estimates under conditions where the assumptions of the Rasch model were violated (i.e., 2PL unidimensional and multidimensional data). Besides the null DIF condition, items with moderate to large DIF were also simulated to evaluate the impact of including/excluding the studied item on DIF detection power.

Method

Two conditions were simulated and studied, one for data that assumed the 2PL model and the other for data that assumed a multidimensional IRT (MIRT) model. The Rasch model is a restricted model, since the discrimination parameter is assumed to be the same for all items in a test. In realistic testing situations, different items often have different discrimination parameters. The 2PL IRT model is usually a more reasonable model that fits the data better. Multidimensionality in the data is another possibility that can be found in real testing situations. The items on a test could measure different constructs, and if the studied item were measuring something different than the items that constitute the matching variable, it might matter whether the studied item was included or not.

For the first type of violation of the Rasch model, the 2PL IRT model was used to simulate data. The 2PL model can be expressed using the following formula:

$$P(U_i = 1 | \theta) = \frac{1}{1 + e^{-1.7 a_i (\theta - b_i)}} , \tag{1}$$

where U_i is the response to item *i*, a_i is the discrimination parameter for item *i*, and b_i is the difficulty parameter for item *i*. The following factors were manipulated: the amount of DIF in the studied item and the difference in the focal group ability distribution from the reference group ability distribution. Since the impact of these factors is unknown, we set the levels of the factors in a more exploratory fashion. Each factor was manipulated to have three levels. Studied items were assigned null (*d* = 0), moderate (*d* = 0.25), and large DIF (*d* = 0.50) in the simulation where *d* was introduced only by differences in the difficulty parameter (*b*). The reference group ability distribution was generated to be normally distributed with a mean of 0 and a standard deviation (SD) of 1. The focal group ability distributions (θ) were generated to be normally distributed with means and SDs of (0, 1), (-0.5, 1), and (-1, 1) to create null, large, and very large group ability differences.

A test of 40 items was generated with 38 items having no DIF across all conditions to be used as the matching criterion and two items being the studied items. Item parameters were drawn from a real testing program of Grade 11 basic math skills. For the reference group, the item parameters remained the same across all studied conditions. For the focal group, the item parameters remained the same for the null DIF condition. For the DIF conditions, the item parameters of the 38 non-DIF items remained the same, while the *b* parameters of the two studied items were changed by *d* to introduce DIF. The direction of DIF was varied for the two studied items. For item 39, *d* was set to be positive to simulate items favoring the reference group. Thus, for the moderate and large DIF conditions, the *b* parameter for item 39 was increased by 0.25 and 0.50, respectively. For item 40, *d* was set to be negative to simulate items favoring the focal group. Thus, for the moderate and large DIF conditions, the *b* parameter for item 40 was decreased by 0.25 and 0.50, respectively. The means and SDs of the item parameters used for simulation across the three DIF conditions are presented in Table 1 for only the focal group. The mean and SD of item parameters used for simulating the reference group responses remained the same as those of the focal group in the null DIF condition. The means and SDs of the *a* parameters for the focal group remained the same across all DIF conditions, because DIF was simulated to be introduced only by differences in the *b* parameters. The means of the *b* parameters for the focal group also remained the same across all DIF conditions because only the *b* parameters for the two studied items changed due to DIF, and the changes canceled each other out ($b_{39f} = b_{39r} + d$ and $b_{40f} = b_{40r} – d$). Thus, only the SDs of the *b* parameters for the focal group were different across the three DIF conditions, due to changes in the *b* parameters.

The nine studied conditions (3 magnitude of DIF levels × 3 group ability difference levels) were replicated 200 times each with two sample sizes (500 and 1,000) in both the focal and reference groups. Table 2 is a summary of the simulation conditions set up for the first type of violation of the Rasch model—the 2PL condition. For the null DIF conditions, since the two studied items were simulated to have no DIF, they could actually be added to the matching criterion with the 38 non-DIF items. Thus, when the studied items were included in the matching criterion in the DIF analyses, all 40 items constituted the matching criterion. When the studied items were excluded in the matching criterion, 39 items (all items except the studied item) constituted the matching criterion. For the moderate and large DIF conditions, the two studied items should not be included in the matching criterion if we assume purification of the criterion was done first. Thus, when the studied items were included in the matching criterion, 39 items (38 non-DIF items plus the studied item) constituted the matching criterion. When the studied items were excluded in the matching criterion, the 38 non-DIF items constituted the matching criterion.

Table 1

Means and Standard Deviations of Item Parameters Used for Simulation With the Two-Parameter Logistic (2PL) Item Response Theory (IRT) Model for the Focal Group

DIF size	Statistics	*A*	*b*
Null DIF	Mean	0.97	-0.19
	SD	0.44	0.60
Moderate DIF	Mean	0.97	-0.19
	SD	0.44	0.61
Large DIF	Mean	0.97	-0.19
	SD	0.44	0.64

Table 2

Factors Varied in the Two-Parameter Logistic (2PL) Item Response Theory (IRT) Model Simulation

	Magnitude of differential item functioning (DIF)		
Matching criterion	Null (*d* = 0)	Moderate (*d* = 0.25)	Large (*d* = 0.50)
Number of items in the matching criterion (include studied item)	40	39	39
Number of items in the matching criterion (exclude studied item)	39	38	38
Group ability difference [Ref: θ~(0, 1)]	Null Foc: θ~(0, 1)	Large Foc: θ~(-0.5, 1)	Very large Foc: θ~(-1, 1)

For the second type of violation of the Rasch model, two-dimensional data were simulated using the 2PL MIRT model. Data were simulated to mimic tests that primarily measure one dimension with a special subset of items measuring an additional dimension. This could happen, for example, if a math test has a small section of items, presented in scenarios, which require a certain level of reading ability to help solve the problems. These specific items could be the primary source of DIF, and how inclusion of the studied item could impact DIF estimation is important to investigate. Data were simulated with the compensatory 2PL MIRT model:

$$P_i[U_i = 1 | (\theta_1, \theta_2)] = \frac{1}{1 + e^{-1.7(a_{i1}\theta_1 + a_{i2}\theta_2 + D_i)}}, \qquad (2)$$

where U_i is the response to item *i*, (θ_1, θ_2) are the examinee abilities on the two dimensions, (a_1, a_2) is the item discrimination parameters on the two dimensions, and D_i is the item difficulty parameter. Similar simulation schemas were used for the multidimensional case. The magnitude of DIF and the group ability difference were simulated to have three levels for each factor. The three magnitudes of DIF were null (*d* = 0), moderate (*d* = 0.25) and large (*d* = 0.50). The reference group ability distribution was simulated to be a bivariate normal distribution with a mean and SD of (0, 1) for both dimensions. The focal group ability distributions were simulated to be bivariate normal distributions with three different means and SDs ([0, 1], [-0.5, 1], and [-1, 1] for both dimensions) to create the three ability difference conditions.

A test of 45 items was simulated with 42 non-DIF items and three studied items. The 42 non-DIF items were simulated with 36 items measuring only the first dimension and six items measuring both dimensions. The three studied items were simulated to measure both dimensions. Item parameters were obtained from the same basic math skills test. For the nine items (six non-DIF and three studied items) that measure both dimensions, the a_2 parameters were set to be the same values as the a_1 values that were obtained from the real test. For the reference group, the item parameters remained the same across all studied conditions. For the focal group, the item parameters remained the same for the null DIF condition. For the DIF conditions, the item parameters of the 42 non-DIF items remained the same, while the *D* parameters of the three studied items were changed by *d* to introduce DIF. Again, the direction of DIF was varied for the three studied items. For items 43 and 44, *d* was set to be positive, indicating DIF favoring the focal group. For item 45, *d* was set to be negative, indicating DIF favoring the reference group. The means and SDs of the item parameters used for simulation across the three DIF conditions are presented in Table 3 for only the focal group. The mean and SD of item parameters used for simulating the reference group responses remained the same as those of the focal group in the null DIF condition. The means and SDs of the a_1 and a_2 parameters for the focal group again remained the same across all DIF conditions, because DIF was introduced only by differences in the *D* parameters. Thus, only the means and SDs of the *D* parameters for the focal group were different across the three DIF conditions [note that the changes in *D* parameters in the moderate DIF condition (d = 0.25) was too small to change the rounded mean statistics]. The group ability distributions on the two dimensions were assumed to be highly correlated. In this simulation study, a correlation of 0.80 was assigned.

Table 3

Means and Standard Deviations of Item Parameters Used for Simulation With the Two-Parameter Logistic (2PL) Multidimensional Item Response Theory (MIRT) Model for the Focal Group

DIF group	Statistics	a_1	a_2	D
No DIF	Mean	0.93	0.08	-0.22
	SD	0.46	0.17	0.59
Moderate DIF	Mean	0.93	0.08	-0.22
	SD	0.46	0.17	0.60
Large DIF	Mean	0.93	0.08	-0.21
	SD	0.46	0.17	0.61

The nine studied conditions were replicated 200 times each with two sample sizes (500 and 1,000) in both the focal and reference groups. Table 4 is a summary of the simulation conditions set up for the second type of violation of the Rasch model—multidimensionality. For the null DIF condition, since the three studied items were simulated to have no DIF, they could actually be added to the matching criterion with the 42 non-DIF items. Thus, when the studied items were included from the matching criterion in the DIF analyses, all 45 items constituted the matching criterion. When the studied items were excluded in the matching criterion, 44 items (all items except the studied item) constituted the matching criterion. For the moderate and large DIF conditions, the three studied items should not be included in the matching criterion if we assume purification of the criterion was done first. Thus, when the studied items were included in the matching criterion, 43 items (42 non-DIF items plus the studied item) constituted the matching criterion. When the studied items were excluded from the matching criterion, the 42 non-DIF items constituted the matching criterion.

Table 4

Factors Varied in Two-Parameter Logistic (2PL) Multidimensional Item Response Theory (MIRT) Model Simulation

	Magnitude of DIF		
Matching criterion	Null (*d* = 0)	Moderate (*d* = 0.25)	Large (*d* = 0.50)
Number of items in the matching criterion (include studied item)	45	43	43
Number of items in the matching criterion (exclude studied item)	44	42	42
Group ability difference [Ref: θ_1~(0,1); θ_2~(0,1)]	Null Foc: θ_1~(0,1); θ_2~(0,1)	Large Foc: θ_1~(-0.5,1); θ_2~(-0.5,1)	Very large Foc: θ_1~(-1,1); θ_2~(-1,1)

Two DIF procedures were used in this study, MH D-DIF and standardized (STD) P-DIF. The resulting DIF estimates were compared to evaluate the impact of including/excluding the studied item in the matching criterion under different conditions.

MH D-DIF is calculated by transforming the estimate of the constant odds-ratio, α_{MH}, onto the Delta metric. Delta is an item difficulty index transformed from the proportion correct (*p*) by converting *p* to a z-score using the inverse of the normal cumulative function and then linearly transforming the z-score to a metric with a mean of 13 and a SD of 4. The approximate standard error for the MH D-DIF is calculated for testing the significance of the statistics.

$$\alpha_{MH} = \left[\sum_m R_{rm}W_{fm} / N_{tm}\right] / \left[\sum_m R_{fm}W_{rm} / N_{tm}\right]$$

$$MH\ D-DIF = -2.35\ln\left[\alpha_{MH}\right]$$

$$SE(\text{MH D-DIF}) = [2.35 / \sum_m (R_{rm}W_{fm} / N_{tm})]*$$

$$\sqrt{\sum_m [(R_{rm}W_{fm} + \alpha_{MH}W_{rm}R_{fm})(R_{rm} + W_{fm} + \alpha_{MH}(W_{rm} + R_{fm}) / 2N_{tm}^2)]} \tag{3}$$

where *R* is the number of examinees answering an item correctly, *W* is the number of examinees answering an item incorrectly, *N* is the number of examinees in a category, subscript *m* is the score level in the matching criteria, subscript *r* indicates the statistic is for the reference group, subscript *f* indicates the statistic is for the focal group, and subscript *t* indicates the statistic is for the total group. When an item does not show DIF, the MH D-DIF will be zero; when an item favors the focal group, the MH D-DIF will be positive; and when an item favors the reference group, the MH D-DIF will be negative.

Donoghue et al. (1993) showed that under the 2PL model, there is a connection between the MH D-DIF and the IRT item difficulty parameter difference:

$$MH\ D-DIF = -4a(b_f - b_r), \tag{4}$$

where *a* is the item discrimination parameter, and b_f and b_r are the item difficulty parameters for the focal and reference groups. This connection provides an evaluation criterion for the MH D-DIF estimates under the 2PL conditions. This evaluation criterion can be extended to the 2PL MIRT model through unidimensional approximation using the following formulas (Reckase, Ackerman, & Carlson, 1988):

$$a' = \sqrt{(a_{i1}^2 + a_{i2}^2)},$$
$$\text{and } b' = \frac{-D_i}{a'}, \tag{5}$$

where *a'* and *b'* are the unidimensional approximation of the discrimination and difficulty parameters and the other notation follows those in Equation 2. Thus, for the 2PL MIRT model, the equivalence between the MH D-DIF and the IRT item difficulty parameter difference can be obtained by substituting the *a* and *b* parameters in Equation 4 with the *a'* and *b'* parameters in Equation 5:

$$MH\ D-DIF_{MIRT} = -4a'(\frac{-D_f}{a'} - \frac{-D_r}{a'}) = 4(D_f - D_r). \tag{6}$$

STD P-DIF is computed as the difference between the observed performance of the focal group (P_f) and the predicted performance of the reference group (P_f^*) matched in ability to the focal group.

$$STD\,P-DIF = P_f - P_f^*$$

$$= \sum_m N_{fm} P_{fm} / \sum_m N_{fm} - \sum_m N_{fm} P_{rm} / \sum_m N_{fm} \qquad (7)$$

where *N* is the number of examinees in a category, *P* is the proportion of examinees getting an item correct, subscript *m* is the score level in the matching criterion, subscript *r* indicates the statistic is for the reference group, and subscript *f* indicates the statistic is for the focal group. Similar to the MH D-DIF statistics, a STD P-DIF of zero indicates no DIF; a STD P-DIF of positive values indicates DIF favoring the focal group; and a STD P-DIF of negative values indicates DIF favoring the reference group.

Under each studied condition, 200 sets of statistics were calculated for each studied item. An item was flagged as showing DIF if the MH D-DIF was categorized as B or C (absolute value of MH-DIF greater than or equal to 1 and was significantly greater than zero at the 0.05 alpha level). The classification of items as having DIF is operationally defined at ETS as follows: an item with an MH D-DIF value that is not significantly different from 0 and/or is smaller than 1 in absolute value is classified as an A-level DIF item (negligible DIF); an item with an MH DIF value that is significantly different from 0 and greater than 1 in absolute value is classified as a B-level DIF item; an item with an MH DIF value that is significantly different from 1 in absolute value and greater than 1.5 in absolute value is classified as a C-level DIF item. Similarly, an item was flagged as showing DIF if the absolute value of STD P-DIF was greater than or equal to 0.10 (see Dorans & Holland, 1992).

Results and Discussion

This section presents the summary statistics of the DIF estimates across replications and the rate of detection under each simulation condition. Only the results for the sample size of 1,000 are presented because (a) the summary statistics (averaged across 200 replications) of the DIF estimates obtained for the 500 sample size conditions are similar to those obtained for the 1,000 sample size conditions, and (b) large standard deviations across replications due to unstable estimates using the smaller sample size (500) caused the detection rates to fluctuate to the extent that interpretation based on them would be inappropriate.[1]

The First Type of Violation of the Rasch Model—2PL IRT Model

Results for the null DIF conditions and the DIF conditions (moderate and large) are presented separately. With the null DIF conditions, estimation bias and Type I error rate are of concern. On the other hand, in the DIF conditions, estimation bias and power rates are of concern.

Tables 5 and 6 contain the summary statistics of the DIF estimates across replications under the null DIF conditions using MH D-DIF and STD P-DIF for the two studied items simulated with the 2PL IRT model. One more decimal place was preserved for the STD P-DIF, since it is on a smaller scale [the delta scale has a mean of 13 and a SD of 4, while the percent correct (*p*) can only vary from 0 to 1]. Under the null DIF condition, the DIF estimates should be zero, since $d = 0$. Consistent MH D-DIF and STD P-DIF results were found for both studied items. When group abilities were the same, including or excluding the studied item in the matching criterion produced similar average DIF estimates close to zero (differences < 0.01 for MH D-DIF and < 0.002 for STD P-DIF). Including the studied item led to smaller SDs across replications, but the differences were too small to demonstrate any significant impact. When group abilities were different, including the studied item consistently produced less biased DIF estimates (closer to zero) than excluding the studied item. As group ability differences went from large to very large, the improvement of including over excluding the studied item became even more pronounced. When the studied item was included in the matching criterion, the average DIF estimates were similar and close to zero regardless of the group ability difference. However, when the studied item was excluded, as group ability differences increased, increasingly negative average DIF estimates were obtained. This showed that excluding the studied item produced systematic bias indicating DIF favoring the high ability group, in this case, the reference group. This is a finding that is consistent with findings from previous studies (e.g., Donoghue et al., 1993).

Tables 7 and 8 contain the percentage of replications flagged as showing DIF (DIF detection rate) under the null DIF conditions using MH D-DIF and STD P-DIF. Since the studied items were simulated to have no DIF under the null DIF condition, the DIF detection rates indicate Type I error rates of the DIF estimates. Again, including the studied item in the matching criterion consistently produced smaller Type I error rates, especially for item 40 when STD P-DIF was used. When group abilities were the same, including or excluding the studied item both produced Type I error rates smaller than 0.05 (5%). When group abilities were different, including the studied item still consistently produced Type I error rates smaller than 0.05, while excluding the studied item produced Type I error rates larger than 0.05 for item 40 when the STD P-DIF was used.

Table 5

Summary Statistics of MH D-DIF Under the Null Differential Item Functioning (DIF) Conditions for the Two-Parameter Logistic (2PL) Item Response Theory (IRT) Model

Ability difference	Inclusion		Exclusion	
	Mean	SD	Mean	SD
	Item 39 (*a* = 0.90, *b* = 0.65)			
None	0.04	0.25	0.04	0.31
Large	-0.01	0.27	-0.18	0.28
Very large	0.02	0.35	-0.24	0.36
	Item 40 (*a* = 0.83, *b* = – 0.60)			
None	-0.03	0.28	-0.02	0.33
Large	0.00	0.25	-0.17	0.26
Very large	-0.01	0.28	-0.31	0.29

Table 6

Summary Statistics of STD P-DIF Under the Null DIF Conditions for the Two-Parameter Logistic (2PL) Item Response Theory (IRT) Model

Ability difference	Inclusion		Exclusion	
	Mean	SD	Mean	SD
	Item 39 (a = 0.90, *b* = 0.65)			
None	0.006	0.034	0.008	0.041
Large	-0.001	0.029	-0.020	0.032
Very large	0.000	0.029	-0.023	0.033
	Item 40 (*a* = 0.83, *b* = – 0.60)			
None	-0.004	0.039	-0.002	0.046
Large	-0.001	0.038	-0.029	0.043
Very large	-0.002	0.045	-0.054	0.052

Table 7

Differential Item Functioning (DIF) Detection Rate of MH D-DIF Under the Null DIF Conditions for the Two-Parameter Logistic (2PL) Item Response Theory (IRT) Model

Ability difference	Inclusion	Exclusion
Item 39 (*a* = 0.90, *b* = 0.65)		
None	0.00	0.00
Large	0.00	0.00
Very large	1.00	3.00
Item 40 (*a* = 0.83, *b* = – 0.60)		
None	0.00	0.00
Large	0.00	0.50
Very large	0.00	1.50

Table 8

Differential Item Functioning (DIF) Detection Rate of STD P-DIF Under the Null DIF Conditions for the Two-Parameter Logistic (2PL) Item Response Theory (IRT) Model

Ability difference	Inclusion	Exclusion
Item 39 (*a* = 0.90, *b* = 0.65)		
None	0.00	2.00
Large	0.00	0.50
Very large	0.00	2.00
Item 40 (*a* = 0.83, *b* = – 0.60)		
None	0.50	2.50
Large	1.00	6.00
Very large	2.50	20.00

Equation 4 was used to transform the differences in item difficulty parameters, simulated for the moderate and large DIF conditions, onto the scale of the MH D-DIF. These transformed values are the population values for comparing the MH D-DIF estimates obtained under different studied conditions. Table 9 contains the population values for the two simulated DIF items. Tables 10 and 11 contain the summary statistics of the DIF estimates across replications under the DIF conditions for the two studied items using MH D-DIF and STD P-DIF.

When there was no group difference, including or excluding the studied item produced similar average DIF estimates with differences smaller than 0.07 for the MH D-DIF and smaller than 0.01 for the STD P-DIF. The average MH D-DIF estimates were close to the population

Table 9

Criterion Values of MH D-DIF for the Studied Items— $MH\ D-DIF=-4a(b_f-b_r)$

Studied item	DIF amount	Moderate (*d* = 0.25)	Large (*d* = 0.50)
Item 39 (*a* = 0.90, *b* = 0.65, *d* > 0)		-0.90	-1.80
Item 40 (*a* = 0.83, *b* = – 0.60, *d* < 0)		0.83	1.66

Table 10

Summary Statistics of MH D-DIF Under the Differential Item Functioning (DIF) Conditions for the Two-Parameter Logistic (2PL) Item Response Theory (IRT) Model

	Moderate DIF				Large DIF			
Ability difference	Inclusion		Exclusion		Inclusion		Exclusion	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Item 39 (*a* = 0.90, *b* = 0.65)								
None	-0.89	0.33	-0.85	0.25	-1.78	0.37	-1.73	0.27
Large	-0.92	0.34	-0.95	0.28	-1.85	0.38	-1.85	0.30
Very large	-0.95	0.37	-1.08	0.36	-1.88	0.41	-1.99	0.35
Item 40 (*a* = 0.83, *b* = – 0.60)								
None	0.81	0.30	0.79	0.24	1.64	0.34	1.57	0.28
Large	0.77	0.29	0.69	0.26	1.58	0.30	1.48	0.27
Very large	0.75	0.28	0.55	0.25	1.57	0.34	1.34	0.31

values for both items for both DIF conditions. However, when group abilities were different, the differences between the average estimates produced when the studied item was included and when the studied item was excluded became increasingly large. When the studied item was included in the matching criterion, the average DIF estimates were close to the population values. For example, the average DIF estimates were -0.89, -0.92, and -0.95 across the three group ability difference conditions for item 39 when simulated with moderate DIF, which were close to the population value of -0.90. However, when the studied item was excluded, the average DIF estimates became far apart from the population value as the group ability difference increased. This effect is most pronounced for item 40 when simulated with large DIF. With large and very large group ability differences, the average DIF estimates became less than 1.50, which would make the classification of the item drop from a C-DIF item to a B-DIF item. The differences between the average DIF estimates under the two conditions (inclusion/exclusion) went from close to 0 to up to 0.23 for the MH D-DIF estimates and from 0.003 to 0.026 for the STD P-DIF as the group ability differences increased. Despite the direction of DIF in the two studied items, the differences between the average DIF estimates obtained by excluding versus including the studied item were consistently negative when group abilities were different (i.e., $\text{D-DIF}_{exclusion} - \text{D-DIF}_{inclusion} < 0$). This indicated a consistent bias in the same direction as observed in the null DIF conditions, that is, a tendency to produce negative bias confounding DIF estimates with the impact of a high ability reference group.

Table 11

Summary Statistics of STD P-DIF Under the Differential Item Functioning (DIF) Conditions for the Two-Parameter Logistic (2PL) Item Response Theory (IRT) Model

	Moderate DIF				Large DIF			
Ability difference	Inclusion		Exclusion		Inclusion		Exclusion	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Item 39 (*a* = 0.90, *b* = 0.65)								
None	-0.109	0.038	-0.113	0.033	-0.208	0.040	-0.218	0.033
Large	-0.090	0.034	-0.099	0.031	-0.165	0.034	-0.178	0.030
Very large	-0.067	0.028	-0.083	0.029	-0.118	0.027	-0.136	0.026
Item 40 (*a* = 0.83, *b* = – 0.60)								
None	0.109	0.038	0.112	0.034	0.210	0.042	0.211	0.036
Large	0.115	0.044	0.109	0.042	0.231	0.042	0.228	0.040
Very large	0.108	0.044	0.082	0.044	0.233	0.051	0.212	0.051

The negative bias observed when the studied item was excluded from the matching criterion produced differential impact on DIF detection rates for the two studied items with different directions of DIF. Tables 12 and 13 contain the DIF detection rates under the DIF conditions for the two studied items simulated with the 2PL IRT model using MH D-DIF and STD P-DIF. When group abilities were the same, even though the average DIF estimates were close regardless of whether the studied item was included or not, the detection rates across the two settings were different because of differences in SDs of the DIF estimates. Excluding the studied item consistently produced smaller SDs of the DIF estimates, leading to lower detection rates (i.e., less power) when MH D-DIF was used. The smaller SDs when the studied item was excluded did not lead to lower detection rates, but actually lead to higher detection rates when

STD P-DIF was used. This was probably due to random variation rather than anything systematic.

Table 12

Differential Item Functioning (DIF) Detection Rates of MH D-DIF Under the DIF Conditions for the Two-Parameter Logistic (2PL) Item Response Theory (IRT) Model

	Moderate DIF		Large DIF	
Ability difference	Inclusion	Exclusion	Inclusion	Exclusion
Item 39 (a = 0.90, b = 0.65)				
None	33.50	25.00	97.00	99.50
Large	38.50	42.00	99.50	99.50
Very large	46.00	61.00	99.50	100.00
Item 40 (a = 0.83, b = – 0.60)				
None	24.50	17.00	98.00	98.50
Large	22.50	13.50	98.00	97.00
Very large	19.00	2.50	94.50	91.50

Table 13

Differential Item Functioning (DIF) Detection Rates of Standardized P-DIF Under the DIF Conditions for the Two-Parameter Logistic (2PL) Item Response Theory (IRT) Model

	Moderate DIF		Large DIF	
Ability difference	Inclusion	Exclusion	Inclusion	Exclusion
Item 39 (a = 0.90, b = 0.65)				
None	61.00	64.50	99.00	100.00
Large	34.00	46.50	98.00	99.50
Very large	12.00	29.00	76.00	90.50
Item 40 (a = 0.83, b = – 0.60)				
None	56.00	68.00	99.50	99.50
Large	62.50	56.50	100.00	100.00
Very large	60.50	34.50	99.50	99.00

When group abilities were different, similar trends were detected for both the MH D-DIF and the STD P-DIF. The consistent negative bias associated with excluding the studied item resulted in larger negative DIF estimates for item 39 and, therefore, produced consistently higher

DIF detection rates (more power). However, the improved power is an artifact, since it was introduced by the negative bias in DIF estimates when the studied item was excluded. On the contrary, for item 40, the negative bias associated with excluding the studied item resulted in smaller positive DIF estimates and, therefore, produced consistently lower DIF detection rates (less power).

The Second Type of Violation of the Rasch Model—Multidimensionality

It is not uncommon in real testing situations to have a small group of items that measure a second dimension that tests a related but different construct. These items are often more prone to showing DIF because of the nature of the second dimension being measured. For example, a language arts exam could contain a section of reading items that require a certain degree of background knowledge (awareness of the topics in the reading passages). The passages, if not carefully chosen, could introduce item bias because different demographic groups could be provided an advantage because of familiarity with certain topics. When DIF occurs on the second dimension, inclusion or exclusion of the studied item could have significant impact, since most items in the matching criterion measure only the first dimension and only a small section of items measure the second dimension.

Tables 14 and 15 contain the summary statistics of the DIF estimates across replications under the null DIF conditions using MH D-DIF and STD P-DIF for the three studied items simulated with the 2PL MIRT model. Similar patterns of results were found for the multidimensional conditions as for the unidimensional 2PL IRT conditions. Consistent MH D-DIF and STD P-DIF results were found for the three studied items. When group abilities were the same, including or excluding the studied item in the matching criterion produced average DIF estimates of zero with very small standard deviations. When group abilities were different, including the studied item in the matching criterion consistently produced less biased average DIF estimates (relatively closer to zero) than excluding the studied item. As group ability differences went from large to very large, the differences in average DIF estimates between the two conditions increased. As group ability differences increased, a consistent negative bias was detected. However, this effect was much larger when the studied item was excluded than included. Excluding the studied item, again, produced systematic bias indicating DIF favoring the high ability group, the reference group.

Table 14

Summary Statistics of MH D-DIF for the Multidimensional Item Response Theory (MIRT) Model—Null Differential Item Functioning (DIF) Conditions

Ability difference	Inclusion		Exclusion	
	Mean	SD	Mean	SD
	Item 43 (a_1 = 0.57, a_2 = 0.54, D = 0.19)			
None	0.00	0.03	0.00	0.02
Large	-0.16	0.19	-0.26	0.18
Very large	-0.33	0.28	-0.54	0.27
	Item 44 (a_1 = 0.70, a_2 = 0.42, D = 0.24)			
None	0.00	0.03	0.00	0.03
Large	-0.11	0.17	-0.21	0.17
Very large	-0.21	0.26	-0.42	0.26
	Item 45 (a_1 = 0.67, a_2 = 0.59, D = 0.60)			
None	0.00	0.01	0.00	0.02
Large	-0.19	0.19	-0.29	0.18
Very large	-0.37	0.26	-0.59	0.25

Table 15

Summary Statistics of STD P-DIF for the Multidimensional Item Response Theory (MIRT) Model—Null Item Functioning (DIF) Conditions

Ability difference	Inclusion		Exclusion	
	Mean	SD	Mean	SD
	Item 43 (a_1 = 0.57, a_2 = 0.54, D = 0.19)			
None	0.000	0.002	0.000	0.001
Large	-0.021	0.027	-0.039	0.028
Very large	-0.039	0.034	-0.075	0.038
	Item 44 (a_1 = 0.70, a_2 = 0.42, D = 0.24)			
None	0.000	0.003	0.000	0.004
Large	-0.015	0.024	-0.032	0.026
Very large	-0.024	0.033	-0.058	0.036
	Item 45 (a_1 = 0.67, a_2 = 0.59, D = 0.60)			
None	0.000	0.001	0.000	0.001
Large	-0.024	0.026	-0.042	0.027
Very large	-0.046	0.034	-0.087	0.038

Tables 16 and 17 contain the DIF detection rates (Type I error rates, in this case) under the null DIF conditions using MH D-DIF and STD P-DIF. When group abilities were the same, including or excluding the studied item produced the same Type I error rate, zero. When group abilities were different, again including the studied item in the matching criterion consistently produced smaller Type I error rates, especially for the large DIF conditions.

Table 16

Differential Item Functioning (DIF) Detection Rates of MH D-DIF for the Multidimensional Item Response Theory (MIRT) Model—Null DIF Conditions

Ability difference	Inclusion	Exclusion
	Item 43 (a_1 = 0.57, a_2 = 0.54, D = 0.19)	
None	0.00	0.00
Large	0.00	0.00
Very large	0.00	5.00
	Item 44 (a_1 = 0.70, a_2 = 0.42, D = 0.24)	
None	0.00	0.00
Large	0.00	0.00
Very large	0.00	0.50
	Item 45 (a_1 = 0.67, a_2 = 0.59, D = 0.60)	
None	0.00	0.00
Large	0.00	0.00
Very large	1.00	6.00

Table 18 contains the population MH D-DIF values obtained using Equation 6 for the three studied items simulated with different amounts of DIF under the 2PL MIRT model. Tables 19 and 20 contain the summary statistics of the DIF estimates across replications using MH D-DIF and STD P-DIF. When group abilities were the same, including or excluding the studied item produced similar average DIF estimates with differences smaller than 0.04 for MH D-DIF. These average DIF estimates were close to the population values indicating good accuracy of estimation when there was no group ability difference. For the STD P-DIF, a similar trend was found when there was no group ability difference (including or excluding the studied item produced differences of 0.005 to 0.012 in average STD P-DIF estimates). However, when group abilities were different, the differences between the estimates produced under the two settings became increasingly large. The differences went from 0.05 to 0.30 for the average MH D-DIF

Table 17

Differential Item Functioning (DIF) Detection Rates of STD P-DIF for the Multidimensional Item Response Theory (MIRT) Model—Null DIF Conditions

Ability difference	Inclusion	Exclusion
	Item 43 (a_1 = 0.57, a_2 = 0.54, D = 0.19)	
None	0.00	0.00
Large	0.00	1.50
Very large	4.50	22.00
	Item 44 (a_1 = 0.70, a_2 = 0.42, D = 0.24)	
None	0.00	0.00
Large	0.50	0.50
Very large	0.50	12.00
	Item 45 (a_1 = 0.67, a_2 = 0.59, D = 0.60)	
None	0.00	0.00
Large	0.50	2.00
Very large	8.00	33.00

Table 18

Criterion Values of MH D-DIF for the Studied Items— $MH\ D - DIF_{MIRT} = 4(D_f - D_r)$

Studied item	DIF amount	Moderate (d = 0.25)	Large (d = 0.50)
Item 43 (a_1 = 0.57, a_2 = 0.54, D = 0.19, $d > 0$)		1.00	2.00
Item 44 (a_1 = 0.70, a_2 = 0.42, D = 0.24, $d > 0$)		1.00	2.00
Item 45 (a_1 = 0.67, a_2 = 0.59, D = 0.60, $d < 0$)		-1.00	-2.00

estimates and from 0.003 to 0.052 for the average STD P-DIF. As group ability difference increased, even including the studied item produced average DIF estimates far from the population values. For example, for item 43 with moderate DIF and very large group ability difference, the average MH D-DIF was 0.60, which was quite different from the population value of 1.00 (by 0.40). However, excluding the studied item produced much worse estimates when group ability difference increased. For example, for the same item under the same condition, excluding the studied item produced average MH D-DIF of 0.34, an average estimate different from the population value of 1.00 by 0.56. Consistent with the 2PL IRT results, excluding the studied item produced smaller positive values for items favoring the focal group (items 43 and 44) and larger negative values for the item favoring the reference group (item 45)

Table 19

Summary Statistics of MH D-DIF Under the Differential Item Functioning (DIF) Conditions for the Multidimensional Item Response Theory (MIRT) Model

	Moderate DIF				Large DIF			
Ability difference	Inclusion		Exclusion		Inclusion		Exclusion	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Item 43 (a_1 = 0.57, a_2 = 0.54, D = 0.19)								
None	0.97	0.12	0.96	0.11	1.95	0.16	1.91	0.16
Large	0.78	0.17	0.64	0.17	1.73	0.15	1.56	0.14
Very large	0.60	0.28	0.34	0.27	1.55	0.27	1.25	0.26
Item 44 (a_1 = 0.70, a_2 = 0.42, D = 0.24)								
None	1.00	0.12	0.98	0.11	1.98	0.16	1.95	0.16
Large	0.83	0.16	0.69	0.15	1.80	0.14	1.63	0.14
Very large	0.74	0.24	0.47	0.23	1.69	0.24	1.39	0.23
Item 45 (a_1 = 0.67, a_2 = 0.59, D = 0.60)								
None	-0.96	0.13	-0.95	0.13	-1.94	0.17	-1.90	0.16
Large	-1.16	0.21	-1.24	0.20	-2.11	0.22	-2.16	0.21
Very large	-1.36	0.28	-1.56	0.27	-2.31	0.30	-2.47	0.29

Table 20

Summary Statistics of STD P-DIF Under the Differential Item Functioning (DIF) Conditions for the Multidimensional Item Response Theory (MIRT) Model

	Moderate DIF				Large DIF			
Ability difference	Inclusion		Exclusion		Inclusion		Exclusion	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Item 43 (a_1 = 0.57, a_2 = 0.54, D = 0.19)								
None	0.128	0.015	0.134	0.015	0.250	0.020	0.262	0.020
Large	0.107	0.022	0.093	0.023	0.239	0.015	0.229	0.015
Very large	0.071	0.034	0.036	0.038	0.195	0.031	0.163	0.035
Item 44 (a_1 = 0.70, a_2 = 0.42, D = 0.24)								
None	0.126	0.015	0.133	0.016	0.245	0.020	0.257	0.020
Large	0.110	0.022	0.097	0.022	0.242	0.014	0.234	0.016
Very large	0.086	0.030	0.053	0.034	0.207	0.029	0.179	0.033
Item 45 (a_1 = 0.67, a_2 = 0.59, D = 0.60)								
None	-0.116	0.015	-0.121	0.016	-0.238	0.020	-0.249	0.021
Large	-0.150	0.028	-0.174	0.029	-0.270	0.029	-0.297	0.031
Very large	-0.159	0.036	-0.207	0.040	-0.256	0.036	-0.308	0.040

when group abilities were different, demonstrating the same negative bias making the DIF estimates confounded with the impact of the high ability reference group.

Differential impact on DIF detection rates (power) was again identified for items with different directions of DIF. Tables 21 and 22 contain the DIF detection rates under the DIF conditions using MH D-DIF and STD P-DIF. When group abilities were the same, detection rates were similar across the two settings for the MH D-DIF with differences smaller than 4%. For the STD P-DIF, due to sensitivity to smaller changes, again excluding the studied item produced higher detection rates (more power). When group abilities were different, the consistent negative bias found in the DIF estimates introduced by excluding the studied item again inflated the power for item 45 and deflated the power for items 43 and 44.

Table 21

Differential Item Functioning (DIF) Detection Rates of MH D-DIF Under the DIF Conditions for the Multidimensional Item Response Theory (MIRT) Model

Ability difference	Moderate DIF		Large DIF	
	Inclusion	Exclusion	Inclusion	Exclusion
	Item 43 (a_1 = 0.57, a_2 = 0.54, D = 0.19)			
None	40.50	38.00	100.00	100.00
Large	13.00	1.50	100.00	100.00
Very large	9.00	1.50	99.00	82.50
	Item 44 (a_1 = 0.70, a_2 = 0.42, D = 0.24)			
None	46.50	44.50	100.00	100.00
Large	12.50	1.50	100.00	100.00
Very large	11.50	0.50	99.50	95.50
	Item 45 (a_1 = 0.67, a_2 = 0.59, D = 0.60)			
None	37.00	33.00	100.00	100.00
Large	76.50	88.50	100.00	100.00
Very large	91.00	99.00	100.00	100.00

Conclusion and Future Directions

Based on simulation results from this study, we can conclude that even when the assumptions of the Rasch model are violated, including the studied item in the matching criterion still reduces the systematic bias in DIF estimation when group abilities are different. For the 2PL MIRT model, including the studied item is not as effective in reducing DIF estimation bias as for the unidimensional 2PL IRT model, especially when the studied items contain DIF. However, the benefits of including the studied item for complex models, such as the 2PL IRT and 2PL MIRT models, are still supported by less biased DIF estimates and more appropriate Type I error rates. For the MH procedure, the bias introduced by excluding the studied item is more pronounced than for the standardization procedure. Thus, the benefit of including the studied item should always be considered when using the MH procedure to evaluate DIF.

Table 22

Differential Item Functioning (DIF) Detection Rates of STD P-DIF Under the DIF Conditions for the Multidimensional Item Response Theory (MIRT) Model

Ability difference	Moderate DIF		Large DIF	
	Inclusion	Exclusion	Inclusion	Exclusion
	Item 43 (a_1 = 0.57, a_2 = 0.54, D = 0.19)			
None	95.50	97.50	100.00	100.00
Large	62.50	36.50	100.00	100.00
Very large	17.00	6.00	100.00	95.50
	Item 44 (a_1 = 0.70, a_2 = 0.42, D = 0.24)			
None	95.50	98.00	100.00	100.00
Large	73.00	45.50	100.00	100.00
Very large	33.00	9.00	99.50	99.50
	Item 45 (a_1 = 0.67, a_2 = 0.59, D = 0.60)			
None	86.00	93.00	100.00	100.00
Large	96.50	99.00	100.00	100.00
Very large	95.50	100.00	100.00	100.00

This study only investigated DIF under the 2PL model where DIF was caused by differential *b* parameters. When both the *a* and *b* parameters were different for the focal group from those for the reference group in a DIF item, it might produce differential impact on DIF estimates when the studied item is included or excluded in the matching criterion. The studied items with different amounts of DIF were average difficulty items with relatively high discrimination parameters. This also limits the generalizability of the results for items with different characteristics (e.g., items with lower discrimination parameters). Test length is another factor that could have an impact on DIF estimates under the two item inclusion conditions.

Whether the benefit of including the studied item in the matching criterion would still exist for longer tests with more than 45 items should be investigated.

This study investigated one specific condition of multidimensionality, that of dimensions caused by specific types of questions in a test. There are many tests that are inherently multidimensional, measuring multiple constructs due to the diverse nature of the overall construct. The correlations between the dimensions for these tests are not necessarily as high as 0.80. DIF could occur on items measuring any one of the dimensions or a combination of different dimensions. How item inclusion in the matching criterion would impact DIF estimation in these situations is of interest too.

Another interesting type of testing situation where item inclusion could have an impact on DIF estimation is when a test is formula-scored. What type of matching criterion to use, a criterion that is rights- or formula-scored, and whether the studied item should be included or not still need to be studied to fully understand the impact of these conditions on operational DIF analyses. As testing programs strive to produce fair and valid assessments, investigations like these will help us improve our DIF analyses procedures and our understanding of the processes.

References

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., & Holland, P. W. (1992). *DIF detection and description: Mantel-Haenszel and standardization* (ETS Research Rep. No. RR-92-10). Princeton, NJ: ETS.

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.

Hamilton, L. S. (1999). Detecting gender-based differential item functioning on a constructed-response science test. *Applied Measurement in Education, 12,* 211-235.

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

Lewis, C. (1993). A note on the value of including the studied item in the test score when analyzing test items for DIF. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 317-319). Hillsdale, NJ: Erlbaum.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

Reckase, M. D., Ackerman, T. A., & Carlson, J. E. (1988). Building a unidimensional test using multidimensional items. *Journal of Educational Measurement, 25*, 193-203.

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.

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.

Zenisky, A. L., Hambleton, R. K., & Robin, F. (2003). Detection of differential item functioning in large-scale state assessments: A study evaluating a two-stage approach. *Educational and Psychological Measurement, 63*, 51-64.

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., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30*, 233-251.

Notes

[1] For the null DIF conditions, due to unstable estimates, the detection rates for conditions with no group ability difference were larger than those for the conditions with group ability differences. The full set of results for the conditions with a sample size of 500 can be acquired from the first author.