# Exploring Crossing Differential Item Functioning by Gender in Mathematics Assessment

**Article** *in* International Journal of Testing · October 2015

DOI: 10.1080/15305058.2015.1057639

**3 authors:**

Yoke Mooi Ong

**3** PUBLICATIONS **27** CITATIONS

SEE PROFILE

Julian Williams
The University of Manchester

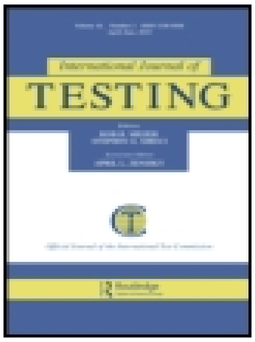
**171** PUBLICATIONS **2,616** CITATIONS

SEE PROFILE

Iasonas Lamprianou
University of Cyprus

**103** PUBLICATIONS **944** CITATIONS

SEE PROFILE



International Journal of Testing

Routledge
Taylor & Francis Group

ISSN: 1530-5058 (Print) 1532-7574 (Online) Journal homepage: http://www.tandfonline.com/loi/hijt20

# Exploring Crossing Differential Item Functioning by Gender in Mathematics Assessment

Yoke Mooi Ong, Julian Williams & Iasonas Lamprianou

**To cite this article:** Yoke Mooi Ong, Julian Williams & Iasonas Lamprianou (2015): Exploring Crossing Differential Item Functioning by Gender in Mathematics Assessment, International Journal of Testing, DOI: 10.1080/15305058.2015.1057639

**To link to this article:** http://dx.doi.org/10.1080/15305058.2015.1057639

Published online: 18 Sep 2015.

*International Journal of Testing*, 00: 1–19, 2015
Copyright © Taylor & Francis Group, LLC
ISSN: 1530-5058 print / 1532-7574 online
DOI: 10.1080/15305058.2015.1057639

# Exploring Crossing Differential Item Functioning by Gender in Mathematics Assessment

Yoke Mooi Ong
*Institute of Teacher Education, Ipoh Campus, Malaysia*

Julian Williams and Iasonas Lamprianou
*School of Environment, Education and Development, University of Manchester, Manchester, United Kingdom*

The purpose of this article is to explore crossing differential item functioning (DIF) in a test drawn from a national examination of mathematics for 11-year-old pupils in England. An empirical dataset was analyzed to explore DIF by gender in a mathematics assessment. A two-step process involving the logistic regression (LR) procedure for detecting uniform and nonuniform DIF was applied to identify crossing DIF. The results showed 36 uniform and 19 nonuniform statistically significant gender DIF items. Out of the 19 nonuniform DIF items, 10 items were crossing DIF. We explained nonuniform DIF using the crossing point in item characteristic curves and the LR-DIF coding scheme. This study showed that crossing DIF exists in empirical data and the findings from this study provide a potentially valuable contribution in understanding such items.

*Keywords:* *crossing differential item functioning, gender, mathematics assessment*

Educational tests are routinely used for diverse and high-stakes purposes, such as selection, placement, and diagnosis for purposes of remedial teaching or certification. The way in which test scores are used to make inferences about examinees' performance is important to fairness and validity (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). Investigating differential item functioning (DIF) is important, especially in the case of high-stakes assessments, because

Correspondence should be sent to Yoke Mooi Ong, Institute of Teacher Education, Ipoh Campus, 31150 Hulu Kinta, Perak, Malaysia. E-mail: oyokemooi@hotmail.com

DIF has been linked to test bias and unfairness. Technically speaking, DIF is said to exist when an item functions differentially for different groups who have the same underlying “ability” on the measured construct (Holland &Thayer, 1988).

In previous work we examined the validity issues arising from DIF in a high-stakes context and more specifically in the context of national testing in English primary schools (Ong, Williams, & Lamprianou, 2011, 2013). In this article, we examine the special case of crossing DIF, which can be difficult to detect when using more traditional DIF techniques designed to measure uniform DIF. The identification of crossing DIF is also very important for the rapidly growing body of international comparative research such as that done by the Trends in International Mathematics and Science Study (TIMMS), the Programme for International Student Assessment (PISA), and the like. Various techniques, such as confirmatory factor analysis (CFA) and item response theory (IRT), are often used to demonstrate that a measurement instrument such as a questionnaire or a test yields comparable results across nations, educational systems, or cultures (for a relevant example of CFA, see Ercikan & Koh, 2005; for a relevant example of IRT, see Meade, Lautenschlager, & Hecht, 2005).

In addition to the methods mentioned, DIF has also been used as a tool to investigate the comparability of test results across cultures. For example, Klieme and Baumert (2001) used DIF analyses to identify the effects of national culture in mathematics education in TIMMS data while Budgell, Raju, and Quartetti (1995) used DIF analyses to study the translation of assessment instruments. The significance of crossing DIF for such international comparisons, then, is that real biases in scores affecting high-profile league tables (i.e., to compare students’ performance ranked by country) and so on might be hidden if studies rely on traditional methods that fail to detect such crossing DIF. DIF has also been used in other disciplines, such as epidemiology, to evaluate the comparability and quality of translations (see, e.g., Bjorner, Kreiner, Ware, Damsgaard, & Bech, 1998). Unfortunately, there is a lack of studies investigating crossing DIF. The recent attention to DIF in international comparisons has debated the relevance and impact of country DIF (e.g., gender DIF across countries) on results such as league tables (e.g., Kreiner & Christensen, 2014). Yet neither this article, nor as far as we know the international studies such as TIMSS and PISA, have even considered the relevance of crossing DIF. Thus even in cases of such validations where it is claimed that DIF has been considered, one might want to ask whether crossing DIF was involved. Since we know that crossing DIF can detect forms of DIF that cannot be detected using more traditional methods, our study becomes even more important since it can be used as a detailed guide for the application of crossing DIF to international comparability studies.

Normally, DIF is examined by comparing the item responses of two groups of examinees often labeled the reference and focal groups. The reference group is expected to be favored by the item, whereas the focal group is expected to be disadvantaged by the item. Several authors have defined uniform and nonuniform DIF (Mellenbergh, 1982; Li & Stout, 1996; Swaminathan & Rogers, 1990). According to Mellenbergh (1982) uniform DIF exists when there is no interaction between ability level and group membership. The probability of a correct response is consistently greater for one group in comparison to another group at all ability levels. On the other hand, nonuniform DIF occurs when there is an interaction between ability level and group membership. The difference in the probabilities of answering an item correctly is not the same across all ability levels. Swaminathan and Rogers (1990) discussed two different types of nonuniform DIF. Both of these types were explained using item characteristic curves (ICCs), which relate the probability of a correct answer to ability level. A disordinal interaction occurs when the ICCs cross in the middle of the ability range. An ordinal interaction occurs when the ICCs cross at either the low end or high end of the ability range. In this case, nonuniform DIF may appear to be uniform DIF across most of the ability range. Li and Stout (1996) used different language to explain uniform and nonuniform DIF. For them, unidirectional DIF includes uniform and nonuniform DIF. They suggested that with nonuniform DIF, the crossing point is far away from the middle of the ability range, but for the crossing DIF, the ICCs cross relatively near the middle of the ability range.

In IRT methodology, the interaction between ability level and the probability of a correct response can be modeled as a difference in the discrimination parameters for two groups. This means that although the difficulty parameter (*b*) is the same (or very similar) for both groups, the discrimination parameter (*a*) is different for the groups. The process results in ICCs that cross. DIF in the IRT sense is conceptualized as the difference in the area between the ICCs for the two groups (Raju, 1988). Uniform and nonuniform DIF are characterized by two parallel and two nonparallel ICCs, respectively.

Examples of ICCs for items with uniform and crossing (disordinal) DIF by gender are shown in Figure 1a and Figure 1b, whereas ICCs for items with unidirectional (ordinal) DIF by gender are shown in Figure 1c and Figure 1d. As can be seen from Figure 1a, there is no interaction between ability level and gender. Figure 1b shows a symmetrical crossing of the two ICCs that occurs at the middle ability level where the probability of a correct response is 0.5. In other words, the magnitude of the area under the curves for the groups in this case may be cancelled out (Penfield, 2003). In general, however, the ICCs of nonuniform DIF may cross at lower or upper ability values (see Figure 1c and 1d). In these cases, the area under the curves for the two groups may not cancel out and actually show a uniform DIF effect. Similarly, for non-IRT DIF detection approaches (e.g., standardization approach; Dorans & Kulick, 1986), the magnitude of a DIF effect is computed by calculating the average difference in probabilities across all ability levels. If methods sensitive only to uniform DIF are used in these scenarios, the nonuniform DIF effect of items with the crossing point near the middle of the

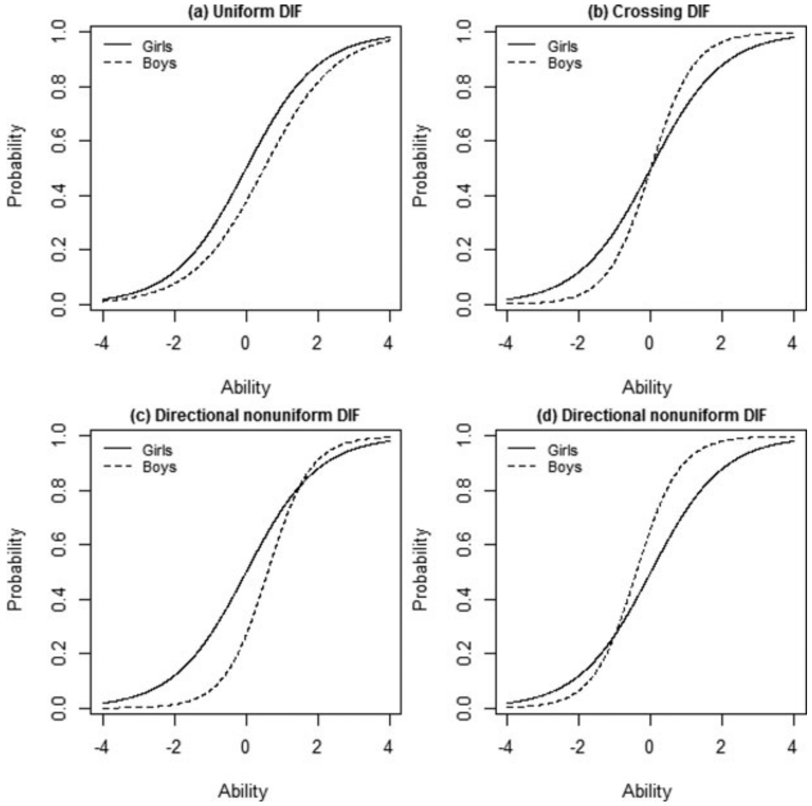


FIGURE 1
Item characteristic curves of uniform and nonuniform differential item functioning.
*Note.* Figures 1c and 1d are noncrossing DIF

ability level is not likely to be detected as statistically significant. For the nonuniform DIF effect with a crossing point significantly higher or lower than the middle of the ability level, the DIF effect may still be detected as statistically significant (Dorans & Kulick, 1986; Mantel & Haenszel, 1959; Rasch, 1960; Shealy & Stout, 1993). Exploring crossing DIF in items is important to validity because these items are hidden from uniform DIF detection methods. In this study, we used Li and Stout's (1996) terminology to differentiate between two types of nonuniform DIF (i.e., unidirectional nonuniform DIF and crossing DIF).

Much research has been conducted to develop statistical methods to detect DIF (Holland & Wainer, 1993; Millsap & Everson, 1993; Penfield & Camilli, 2007; Potenza & Dorans, 1995). In particular, crossing-SIBTEST (Li & Stout, 1996) has been developed specifically to detect crossing DIF whereas the logistic regression (LR) procedure developed by Swaminathan and Rogers (1990) can identify both uniform and nonuniform DIF.

Research investigating the causes of DIF have viewed DIF effects from a multidimensional perspective, where item responses are modeled by at least one secondary dimension in addition to the primary dimension (Ackerman, 1992; Roussos & Stout, 1996; Shealy & Stout, 1993). The presence of DIF may indicate unintended item-level multidimensionality (Ackerman, 1992; Camilli, 1992). Several researchers have attempted to explain why DIF occurs through their examination of bundles of items that have common characteristics (see Gierl, Bisanz, Bisanz, & Boughton, 2003; Mendes-Barnett & Ercikan, 2006; Ryan & Chiu, 2001; Walker & Beretvas, 2001; Walker, Zhang, & Sauber, 2008). The process of testing bundles of items for differences is known as differential bundle functioning (DBF). DBF is said to exist when bundles of items with common characteristics function differentially for different groups with the same ability (Douglas, Roussos, & Stout, 1996). The performance differences (between bundles) by groups may be the result of construct-relevant variance (leading to benign DBF) or construct-irrelevant variance (leading to adverse DBF). Benign DBF occurs when a bundle of items that measures the achievement construct of interest and an additional secondary dimension that *was* also intended to be measured by the bundle (e.g., a bundle of mathematics items that require relevant mathematical vocabulary knowledge). In contrast, adverse DBF occurs when a bundle of items that measures the achievement construct of interest and an additional secondary dimension that was not intended to be measured (e.g., a bundle of items from a reading passage about football).

In our previous work (Ong et al., 2011) using a subset of this dataset, we hypothesized sources of DIF based on the review of literature. Our goal was to predict and explain DIF by gender in a mathematics test via DBF. The results suggested that bundles that measured *mental test, using and applying mathematics*, and *redundant illustrations* were relatively easier for boys than for girls. By contrast, *calculation* and *high language demand* bundles were relatively easier for girls than for boys. In this case, the performance differences by gender on *mental test*, *using and applying mathematics*, and *calculation* bundles may have construct-relevant variance because these bundles measured the construct of interest and additional dimensions that were intended to be measured by the items, which resulted in benign DBF. On the other hand, the performance differences by gender on *redundant illustrations* bundle may have construct-irrelevant variance because the bundle measured the construct of interest and an additional dimension that was not intended to be measured, which resulted in adverse DBF. Therefore, it is likely that factors such as language and curriculum domain may represent a relevant secondary dimension in the data, which can explain some gender differences in performance relevant to our discussions later in this article.

A review of the literature on gender DIF in mathematics tests suggested that the item characteristics displaying DIF in favor of boys include: (a) problem solving (Innabi & Dodeen, 2006; Mendes-Barnett & Ercikan, 2006; Ryan & Chiu, 2001), (b) difficult items (Penner, 2003), (c) space and shape (Gierl et al., 2003; Innabi & Dodeen, 2006; Liu & Wilson, 2009; Williams, Hadjidemetriou, Ryan, & Jones, 1999), and (d) mental mathematics (Ong et al., 2011). However, item characteristics exhibiting DIF in favor of girls include: (a) computation (Mendes-Barnett & Ercikan, 2006), (b) algorithms (Innabi & Dodeen, 2006; Williams et al., 1999), (c) verbal demand (Gallagher et al., 2000; Ong et al., 2011; Walker et al., 2008; Williams et al., 1999), and (d) memorization (Gierl et al., 2003). Past research on nonuniform DIF by gender in mathematics assessments rarely investigate the causes of nonuniform DIF but instead compare the power of detecting nonuniform DIF using different DIF detection methods (Abedlaziz, Wail, & Zaharah, 2011; Gierl, Khaliq, & Boughton, 1999).

The literature on gender DIF in mathematics tests draws mostly from uniform DIF. The LR procedure can detect both uniform and nonuniform DIF. However, when the null hypothesis of nonuniform DIF is rejected, suggesting the presence of DIF, it is not clear which group is favored and at what ability level. Therefore, the main objectives of this study were to: (a) identify different types of nonuniform gender DIF in test items, (b) determine the crossing point of different ICCs of nonuniform gender DIF using the estimated LR parameters, (c) interpret nonuniform gender DIF by classifying it using a LR-DIF coding scheme, and (d) investigate the effect of crossing DIF on the measurement of examinees' test performance.

## METHODS

### Instruments

As described in our previous work (Ong et al., 2011, 2013), the test consisted of three subtests administered separately, namely Test A (Qualification and Curriculum Authority, 2007a), Test B (Qualification and Curriculum Authority, 2007b), and Mental Mathematics (Qualification and Curriculum Authority, 2007c). The total score was 100 points. Test A was a standard pencil-and-paper test and worth 40 points. Test B allowed pupils to use calculators and was worth 40 points. The Mental Mathematics test required pupils to respond in writing to items that were read out to them within time constraints and was worth 20 points. All of these items used an open-ended response format, consisting of a combination of calculation, problem solving, and mathematical reasoning, which were scored dichotomously. Table 1 shows the number of items at each curriculum domain on each subtest.

TABLE 1
Distribution of Items According to Curriculum Domains and Subtests

| | Number of Items | | | |
|---|---|---|---|---|
| Curriculum Domains | Test A | Test B | Mental | Total |
| Using and applying mathematics | 7 | 8 | — | 15 |
| Numbers and number system | 15 | 16 | 17 | 48 |
| Shape, space and measures | 11 | 8 | 3 | 22 |
| Handling data | 7 | 8 | — | 15 |
| Total | 40 | 40 | 20 | 100 |

### Examinees

The data set was drawn from a national examination of mathematics for 11-year-old primary school pupils in England. The sample consisted of "live" and pre-test examinees. The live sample consisted of 1029 boys (51%) and 971 girls (49%) who took the national examination at the end of primary schooling. The pre-test sample consisted of 455 boys (51%) and 436 girls (49%) who took the pre-test a year earlier. The mean total score for boys ($M = 68.82$, $SD = 21.33$) was higher than that for girls ($M = 64.74$, $SD = 21.65$).

## DATA ANALYSIS

The LR procedure (Swaminathan & Rogers, 1990) was used to detect DIF, which involves modeling the probability of a person responding correctly to an item as a function of $X$ (total observed score), $G$ (group membership, in our case girls and boys are dummy coded 0 for boys and 1 for girls), and $XG$ (interaction of total observed score and group membership). The LR equation can be written as:

$$P(Y = 1|X, G) = \frac{e^z}{1 + e^z}, \tag{1}$$

$$\text{where } z = b_0 + b_1 X + b_2 G + b_3 (XG).$$

The variable $Y$ is the examinee's item response score coded as 1 (correct) or 0 (incorrect), the regression parameters $b_0$, $b_1$, $b_2$, and $b_3$ represent the intercept and weights for total score, group membership, and total score by group interaction terms, respectively. In other words, is the difficulty of the items (or easiness) for the reference category (boys) and $b_0 + b_2$ for girls. Also, $b_1$ is the slope/discrimination for boys and $b_1 + b_3$ for girls.

To test the null hypothesis concerning DIF using the LR procedure, the likelihood ratio test was conducted using a series of three nested models as described by Penfield and Camilli (2007). The simplest model, Model 1, does not consider any form of DIF. Model 1 represents the probability of a correct response only as a function of total observed score ($X$).

$$\text{Model 1} \ : \ z = b_0 + b_1 X \tag{2}$$

Model 2 expresses the probability of a correct response as a function of total observed score ($X$) and group membership ($G$) and thus contains a term ($b_2$) that is associated with uniform DIF.

$$\text{Model 2} \ : \ z = b_0 + b_1 X + b_2 G \tag{3}$$

Model 3 represents the probability of a correct response as a function of total observed score ($X$), group membership ($G$), and an interaction term between total observed score and group membership ($XG$). This model contains the terms $b_2$ and $b_3$ that are associated with uniform and nonuniform DIF.

$$\text{Model 3} \ : \ z = b_0 + b_1 X + b_2 G + b_3 (XG) \tag{4}$$

A statistical test for nonuniform DIF was assessed using a likelihood ratio test by comparing the likelihood of Model 3 with the likelihood of Model 2. The resulting statistic follows a chi-square distribution with one degree of freedom. When $b_3 = 0$, this suggested that nonuniform DIF did not exist (or was insignificant). In this situation, the presence of uniform DIF was assessed using a likelihood ratio test by comparing the likelihood of Model 2 with the likelihood of Model 1. The resulting statistic also follows a chi-square distribution with one degree of freedom. An insignificant result occurred when $b_2 = 0$, indicating that no significant uniform DIF existed. The LR procedure was analyzed using the R software and the parameters were estimated using a *glm* function in the generalized linear model, which is part of the R core program (R Development Core Team, 2011).

Effect size measures are important to consider because they help avoid statistically significant results that are not practically important but can occur due to large sample sizes, regardless of the DIF methods used. The Education Testing Services (ETS) established a DIF classification effect size system based on the Mantel-Haenszel DIF (MH DIF) procedure. The method uses a three-category classification system for identifying uniform DIF. The categories are negligible, moderate, and large (Dorans & Holland, 1993). Penfield and Camilli (2007) demonstrated that for the LR procedure, the uniform DIF parameter $b_2$ can be interpreted as a common log-odds ratio, analogous to the MH DIF common log-odds ratio. Therefore, the ETS classification framework of MH DIF for uniform DIF is applicable to the LR procedure. However, according to the authors, this equivalence does not exist for nonuniform DIF because the MH DIF method was developed specifically to detect uniform DIF. Others have developed effect size measures for the LR procedure (Jodoin & Gierl, 2001). However, according to Hidalgo and Lopez-Pina (2004), these criteria "based on LR appeared to be insensitive to the specified DIF conditions" (p. 914). In addition, Finch and French (2007) pointed out that there is no established framework for classifying effect size measures for nonuniform DIF regardless of the detection methods used. Although classifications of DIF on the basis of effect size measures are important in any DIF study, the fact that there is no existing established classification framework for nonuniform DIF, we will not classify nonuniform DIF items as negligible, moderate, or large.

To determine the crossing point of the two curves, let us write down the equation of the log odds for the focal and reference groups ($G = 0$ for reference group e.g., boys, G = 1 for focal group e.g., girls):

For the focal group: $\mathrm{z} = b_0 + b_1X + b_2 + b_3(X)$

For the reference group: $\mathrm{z} = b_0 + b_1X$

Let $P_c$ (the crossing point) be the ability score where the two curves cross. This is the point where the probability of answering an item correctly is the same for the focal group and the reference group. In other words, the log odds of answering an item correctly are the same for the focal group and the reference group.

$$b_0 + b_1 P_\mathrm{c} = b_0 + b_1 P_\mathrm{c} + b_2 + b_3 P_\mathrm{c}$$

Solving the equation, $P_c$ equals minus the ratio of $b_2$ and $b_3$.

$$P_\mathrm{c} = -\ b_2/b_3$$

Taking the $P_c$ as the boundary, and the value of the estimated $b_3$ parameters, an item either favors girls or boys in the region below $P_c$ and either favors boys or girls in the region above $P_c$. If an item favors girls in the region below $P_c$, then for the region above $P_c$, it should favor boys and vice-versa. We introduce a coding scheme to explain nonuniform DIF, which includes two signs ("−" and "+") and two letters ("G" for girls and "B" for boys). The region below $P_c$ is labeled with a negative sign ("−") whereas the region above $P_c$ is represented by a positive sign ("+"). To illustrate the LR-DIF coding scheme, the log odds is plotted against ability (i.e., $z$-total score) and by gender as shown in Figure 2. In this case, for the

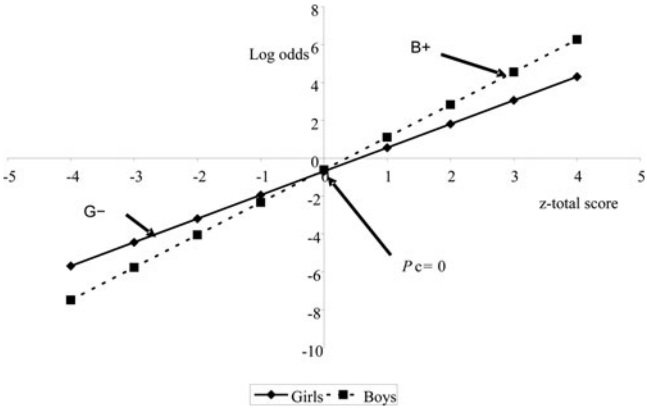


FIGURE 2
Log odds for an item coded G−B+ with $b_3 > 0$ and $P_c = 0$.

region below $P_c$ the item favors girls (G−) whereas for the region above $P_c$ the item favors boys (B+). Combining both the codes for the regions below and above $P_c$, we have the code G−B+. The code G−B+ shows that this item favors girls in the region below $P_c$ and favors boys in the region above $P_c$ whereas the reverse effects is true for the code B−G+.

We compared the results of nonuniform DIF identified by the LR procedure in Model 3 with those of uniform DIF identified by the LR procedure in Model 2. We then categorized these flagged items into probable crossing and unidirectional nonuniform DIF. To evaluate the effect of crossing DIF on the measurement of examinees' test performance, these crossing DIF items were then deleted from the test. We calculated examinees' gain score, which is the difference of examinees' score before and after the deletion of crossing DIF. The nonparametric Mann-Whitney $U$ test was used to compare mean rank gain score distributions of boys and girls for the low ability and high ability groups for combined, live, and pre-test samples.

## RESULTS

There were 36 items with statistically significant uniform DIF effects detected using the LR procedure (18 favored girls and 18 favored boys), with effect size measure based on the absolute $b_2$ estimated parameters ranging from 0.21 logits to 0.75 logits. Using the ETS MH log-odds ratio classification framework, there were 24 negligible, 11 moderate, and 1 large uniform DIF items. We did not expect to see much statistically significant uniform DIF in this test, because the test is

TABLE 2
Results of DIF Items by Gender

| | Uniform DIF Model 2 | Nonuniform DIF Model 3 | | | |
|---|---|---|---|---|---|
| Item Name | $b_2$ | $b_2$ | $b_3$ | $Pc = -b_2/b_3$ | LR Coding |
| **ma10a** | **–0.054** | **−0.235** | **−0.247**$^{*}$ | **−0.95** | **B−G+** |
| **ma20ii** | **–0.137** | **−0.190** | **0.262**$^{*}$ | **0.73** | **G−B+** |
| **ma24** | **0.175** | **−0.004** | **0.542**$^{***}$ | **0.01** | **G−B+** |
| **mb17i** | **–0.099** | **−0.100** | **0.392**$^{**}$ | **0.26** | **G−B+** |
| **mb17ii** | **–0.056** | **−0.085** | **0.356**$^{**}$ | **0.24** | **G−B+** |
| **mb22b** | **0.122** | **0.171** | **−0.252**$^{*}$ | **0.68** | **B−G+** |
| **m3** | **0.133** | **0.365**$^{*}$ | **0.325**$^{*}$ | **−1.12** | **G−B+** |
| **m6** | **–0.131** | **0.025** | **0.275**$^{*}$ | **−0.09** | **G−B+** |
| **m9** | **–0.014** | **0.077** | **0.267**$^{*}$ | **−0.29** | **G−B+** |
| **m17** | **0.138** | **0.358**$^{**}$ | **0.421**$^{**}$ | **−0.85** | **G−B+** |
| ma6bi | –0.436$^{***}$ | −0.873$^{***}$ | −0.508$^{***}$ | −1.72 | B−G+ |
| ma8a | 0.269$^{*}$ | 0.046 | −0.346$^{*}$ | 0.13 | B−G+ |
| ma8b | 0.304$^{**}$ | 0.196 | −0.286$^{*}$ | 0.69 | B−G+ |
| ma15b | –0.316$^{***}$ | −0.355$^{***}$ | 0.298$^{*}$ | 1.19 | G−B+ |
| mb18 | 0.224$^{*}$ | 0.288$^{**}$ | 0.295$^{*}$ | −0.98 | G−B+ |
| mb24 | 0.290$^{**}$ | 0.225$^{*}$ | 0.375$^{**}$ | −0.60 | G−B+ |
| m11 | 0.347$^{***}$ | 0.396$^{***}$ | 0.264$^{*}$ | −1.50 | G−B+ |
| m19 | –0.217$^{*}$ | −0.301$^{**}$ | 0.269$^{*}$ | 1.12 | G−B+ |
| m20 | 0.489$^{***}$ | 0.398$^{***}$ | 0.267$^{*}$ | −1.50 | G−B+ |

$^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

routinely pretested uniform DIF. Since this article focused on nonuniform DIF, these uniform DIF items will not be discussed in further detail.

Table 2 depicts results of DIF by gender using the LR procedure where 19 items exhibited statistically significant nonuniform DIF at the 95% confidence level. Out

TABLE 3
Description of Mental Mathematics Items with Crossing DIF

| Item Name | Description of Item | Comment |
|---|---|---|
| m3 (G−B+) | Multiply ninety-one by two. | Multiplication |
| m6 (G−B+) | Multiply eight by four, then add fifty. | Multiplication and addition |
| m9 (G−B+) | How many twos are there in four hundred and forty? | Division or multiplication |
| m17 (G−B+) | Oranges cost fifteen pence each. I buy four oranges.<br>How much change do I get from a two pound coin? | Multiplication and subtraction |

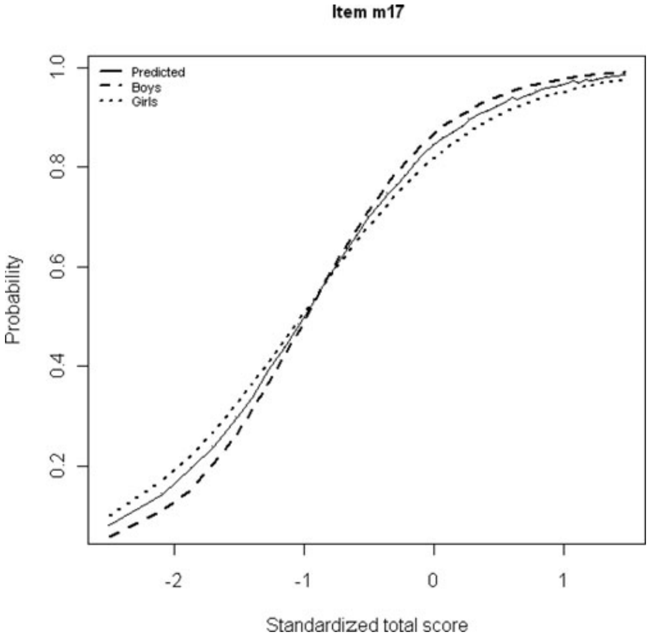


FIGURE 3
Item characteristic curves of a crossing differential item functioning.

of the 19 nonuniform DIF items, 10 items were not detected as uniform DIF. These items (in bold) showed crossing DIF. Eight items were coded as G−B+, indicating that they favored girls in the region below $P_c$ and favored boys in the region above $P_c$.

Four out of the 10 crossing DIF items used mental mathematics as shown in Table 3. These items are coded as G−B+ using the LR DIF coding scheme. Figure 3 shows an example of such an item (m17). Item m17 favors girls at the lower ability levels and favors boys at the higher ability levels. From an IRT perspective, the ICCs cross because the discrimination parameter for item m17 is larger for boys than for girls. There are many factors that may explain why crossing DIF occurs. We speculate on one of the possible causes of crossing DIF here. We hypothesized that solving these items required two steps. First, there is a certain language comprehension threshold required to understand the tasks that favors girls. This is then followed by a mental mathematics execution, which involves problem solving using either multiplication or division that favors boys. In other words, solving these items required at least two skills, namely language comprehension to access the problem and mental computation or problem solving. Hence, the presence of two secondary dimensions may be at work here. Mental mathematics is part of the primary school mathematics curriculum, but language comprehension may be construct-irrelevant to the mathematics items. It is widely known that girls often perform better than boys in language and boys often perform better in mental mathematics than girls (Lynn & Irwing, 2008; Ong et al., 2011). Higher ability boys might have overcome the language comprehension threshold, and their higher competence in mental mathematics, involving the more difficult multiplication or division operation may explain their better performance than the girls.

Two of the items (ma10a and mb22b) are from the handling data curriculum domain, which require examinees to interpret graphs. These items are coded as B−G+ in the LR coding scheme (see Table 2). This means that these items favored boys at the lower ability levels and favored girls at the higher ability levels resulting in discrimination parameters for girls that were larger than for boys. For item mb22b, examinees were presented with a line graph, which showed the height of a candle as it burns and asked "How long does the candle take to burn down from 16 cm to 4 cm?" We hypothesized that solving this item required two subskills, (graphical skills in the interpretation of information represented on the axes and language comprehension to understand the task). Hence, two secondary dimensions may be impacting performance in opposite directions. The literature states that boys outperform girls in interpreting graphs (Lowrie & Diezmannb, 2011) and girls outperform boys in language comprehension (Gallagher et al., 2000; Ong et al., 2011; Walker et al., 2008; Williams et al., 1999). In other words, lower ability boys had an advantage over girls in interpreting the axes of the line graphs. However, higher ability girls may have overcome the threshold of interpreting graphs and, with their higher ability in language comprehension, may explain why they outperformed boys.

Our explanation of crossing DIF is in line with the theory proposed by Penfield (2010) for polytomous data. Indeed, solving these mathematical items required separate steps along with multiple skills, and different groups may have relative advantages on these skills.

All of these arguments are contextual and essentially qualitative. Identifying the presence of secondary dimensions in crossing DIF in an empirical dataset is complicated, and there are many factors that may explain why crossing DIF occurs. Future research should consider similar studies that work to investigate and explain crossing DIF in open-ended dichotomously scored items. Some may argue that crossing DIF might occur by chance and may differ with different samples. Therefore, more research is needed in order to verify whether crossing DIF was local in this study or indeed a more general phenomenon in mathematics testing.

Recall that the sample consisted of live and pretest examinees. Examinees with a $z$-total score equal to or greater than zero were categorized as high ability and those with $z$-total score less than zero were considered low ability. Six groups of examinees (combined high, combined low, live high, live low, pretest high, and pretest low) were formed to compare the effects of deleting ten crossing DIF items

TABLE 4
Boys' and Girls' Mean Ranks Gain Scores[1] of Different Ability Levels

| | Boys | | Girls | | | |
|---|---|---|---|---|---|---|
| Group | *n* | Mean Rank | *n* | Mean Rank | *U* | Effect Size |
| Combined high | 897 | 765 | 735 | 858 | 291772*** | 0.10 |
| Combined low | 587 | 638 | 672 | 620 | 191772 | 0.02 |
| Live high | 639 | 556 | 547 | 625 | 154306*** | 0.10 |
| Live low | 390 | 412 | 424 | 403 | 80829 | 0.02 |
| Pre-test high | 258 | 208 | 188 | 234 | 21497* | 0.10 |
| Pre-test low | 197 | 227 | 248 | 218 | 23477 | 0.03 |

gain scores = total score of 90 items − total score of 100 items.
$^{*}p < .05$, $^{***}p < .001$.

as shown in Table 4. The term combined refers to live and pretest examinees. The gain score was computed by finding the difference between the total score of examinees before and after deleting 10 crossing DIF items. The Kolmogorow-Smirnov statistic was used to assess the normality distribution of gain scores. A significant result ($p = .001$) indicated that the gain score distribution violated the normality assumption. Therefore, a nonparametric Mann-Whitney $U$ test was used to compare differences in the mean rank gain score distributions of boys and girls. The results showed that by deleting 10 crossing DIF items, the girls' mean rank gain scores were higher than the boys at the upper ability levels, and the boys' mean rank gain scores were higher than the girls at the lower ability levels as expected. Three groups, combined high ($U = 291772$, $p < .001$), live high ($U = 154306$, $p < .001$), and pretest high ($U = 21497$, $p = .036$) were statistically significantly different from zero at the .05 level of significance. For the combined high group, the mean rank of girls (*mean rank* = 858) was higher than the mean rank of boys (*mean rank* = 735). As for the live high group, the mean rank of girls (*mean rank* = 625) was also higher than the mean rank of boys (*mean rank* = 556). Similarly, for the pretest high group, the mean rank of girls (*mean rank* = 234) was higher than the mean rank of boys (*mean rank* = 208). The other groups (combined low, live low, and pretest low) were statistically insignificant. The effect size measure ranges from 2% to 10%, with the highest effect size shown at the high ability levels. Often these items are not immediately deleted from the test because doing so may cause the test to be unrepresentative of the curriculum domain. According to Zumbo (1999), the decision whether to remove items with DIF from a test depends on the effect of these items on the measurement of examinees, or more importantly, the severity of the consequences of an error.

## DISCUSSION AND CONCLUSION

In this study, data from a high-stakes national examination were used to explore crossing gender DIF in mathematics tests. Test items used in national examinations have usually undergone pretesting to ensure high quality. However, test developers are more likely to use DIF methods that are sensitive to uniform DIF, such as Mantel Haenszel (Holland & Thayer, 1988; Mantel & Haenszel, 1959), standardization approach (Dorans & Kulick, 1986), SIBTEST (Shealy & Stout, 1993), or Rasch model (Rasch, 1960). It is commonly assumed that nonuniform DIF is rare in practice. This study has shown that nonuniform DIF, in particular crossing DIF, exists in empirical data even in high-stakes testing contexts. In this article, we described a process to detect crossing DIF using the LR procedure. Statistically significant nonuniform DIF items detected by Model 3 in the LR procedure that were identified as statistically insignificant uniform DIF by Model 2 were categorized as crossing DIF.

The distinction between unidirectional nonuniform DIF and crossing DIF is important in order to identify DIF cancellation effects at the item level. This study has shown that unidirectional nonuniform DIF does not cause substantial DIF cancellation effects. However, crossing DIF can cause significant DIF cancellation effects, thus making DIF difficult to detect when using methods that are sensitive only to uniform DIF.

The proposed nonuniform DIF coding scheme of the LR procedure provides a tool for describing nonuniform DIF by identifying the crossing point of the ICCs. The process could help researchers determine which group is favored by the item at each extreme of the ability scale. This coding scheme provides a systematic way of describing nonuniform DIF, which may be useful to DIF researchers and practitioners as they discuss different types of DIF.

Angoff (1993) stated, “It has been reported by test developers that they are often confronted by DIF results that they cannot understand; and no amount of deliberation seems to help explain why some perfectly reasonable items have large DIF values” (p. 19). We were also concerned about crossing DIF items, which are likely to favor low ability girls and favor high ability boys or vice-versa. We speculated that one possible explanation of crossing DIF for mental mathematics and handling data items were the two secondary dimensions that impacted item performance in different directions (i.e., girls’ advantaged on language demand and boys’ advantaged on mental calculation and interpreting graphs). However, we acknowledge that our speculation on the causes of crossing DIF is a qualitative judgment and we do not claim to have conclusively identified these secondary dimensions. This study should be followed up by confirmatory studies to explain crossing DIF in the future.

We have acknowledged that this study had its limitations. Additionally, there is currently no commonly well-accepted system for classifying effect size measures in the nonuniform DIF case. Therefore, future research should develop such a system.

## ACKNOWLEDGMENTS

This article is a part of the first author's doctoral thesis at the University of Manchester. The authors wish to thank Malcolm Hayes and Qualification Curriculum Authority of England for granting access to the dataset used in this study.

## REFERENCES

Abedlaziz, N., Wail, I., & Zaharah, H. (2011). Detecting a gender-related DIF using logistic regression and transformed item difficulty. *US-China Education Review, B* (5), 734–744.

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, *29*(1), 67–91. doi:10.1111/j.1745-3984.1992.tb00368.x.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3–23). Hillsdale, NJ: Lawrence Erlbaum.

Bjorner, J. B., Kreiner, S., Ware, J. E., Damsgaard, M. T., & Bech, P. (1998). Differential item functioning in the Danish translation of the SF-3. *Journal of Clinical Epidemiology*, *51*(11), 1189–1202.

Budgell, G. R., Raju, N. S., & Quartetti, D. A. (1995). Analysis of differential item functioning in translated assessment instruments. *Applied Psychological Measurement*, *19*(4), 309–321. doi:10.1177/014662169501900401.

Camilli, G. (1992). A conceptual analysis of differential item functioning in terms of a multidimensional item response model. *Applied Psychological Measurement*, *16*(2), 129–147. doi:10.1177/014662169201600203.

Department for Education and Employment & Qualifications and Curriculum Authority. (1999). *Mathematics: The national curriculum for England, key stages 1–4*. London: Crown, QCA.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Hillsdale, NJ: Lawrence Erlbaum.

Dorans, N. J., & Kulick, E. M. (1986). Demonstrating the utility of the standardization approach assessing unexpected differential item performance on the scholastic aptitude test. *Journal of Educational Measurement*, *23*(4), 355–368. doi:10.1111/j.1745-3984.1986.tb00255.x.

Douglas, J. A., Roussos, L. A., & Stout, W. (1996). Item-bundle DIF hypothesis testing: Identifying suspect bundles and assessing their differential functioning. *Journal of Educational Measurement*, *33*(4), 465–484. doi:10.1111/j.1745-3984.1996.tb00502.x.

Ercikan, K., & Koh, K. (2005). Examining the construct comparability of the English and French versions of TIMSS. *International Journal of Testing*, *5*(1), 23–35. doi:10.1207/s15327574ijt0501.

Finch, W. H., & French, B. F. (2007). Detection of crossing differential item functioning: A comparison of four methods. *Educational and Psychological Measurement*, *67*(4), 565–582. doi:10.1177/0013164406296975.

Gallagher, A. M., De Lisi, R., Holst, P. C., McGillicuddy-De Lisi, A. V., Morely, M., & Cahalan, C. (2000). Gender differences in advanced mathematical problem solving. *Journal of Experimental Child Psychology*, *75*, 165–190. doi:10.1006/jecp.1999.2532.

Gierl, M. J., Bisanz, J., Bisanz, G. L., & Boughton, K. A. (2003). Identifying content and cognitive skills that produce gender differences in mathematics: A demonstration of the multidimensionality-based DIF analysis paradigm. *Journal of Educational Measurement*, *40*(4), 281–306. doi:10.1111/j.1745-3984.2003.tb01148.x.

Gierl, M. J., Khaliq, S. N., & Boughton, K. (1999, June). *Gender differential item functioning in mathematics and science: Prevalence and policy implications*. Paper presented at the Annual Meeting of the Canadian Society for the Study of Education, Sherbrooke, Québec, Canada.

Hidalgo, M. D., & Lopez-Pina, J. A. (2004). Differential item functioning detection and effect size: A comparison between logistic regression and Mantel-Haenszel procedures. *Educational and Psychological Measurement*, *64*(6), 903–915. doi:10.1177/0013164403261769.

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

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum

Innabi, H., & Dodeen, H. (2006). Content analysis of gender-related differential item functioning TIMMS items in mathematics in Jordan. *School Science and Mathematics*, *106*(8), 328–337. doi:10.1111/j.1949-8594.2006.tb17753.x.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I error and power ratees using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education*, *14*(4), 329–349. doi:10.1207/S15324818AME1404_2.

Klieme, E., & Baumert, J. (2001). Identifying national cultures of mathematics education: Analysis of cognitive demands and differential item functioning in TIMSS. *European Journal of Psychology of Education*, *16*(3), 385–402.

Kreiner, S., & Christensen, K. (2014). Analyses of model fit and robustness. A new look at the PISA scaling model underlying ranking of countries according to reading literacy. *Psychometrika*, *79*(2), 210–231. doi:10.1007/s11336-013-9347-z.

Li, H. H., & Stout, W. (1996). A new procedure for detection of crossing DIF. *Psychometrika*, *61*(4), 647–677.

Liu, O. L., & Wilson, M. (2009). Gender differences in large-scale math assessments: PISA trend 2000 and 2003. *Applied Measurement in Education*, *22*(2), 164–184. doi:10.1080/08957340902754635.

Lowrie, T., & Diezmannb, C. M. (2011). Solving graphics tasks: Gender differences in middle-school students. *Learning and Instruction*, *21*(1), 109–125. doi:10.1016/j.learninstruc.2009.11.005.

Lynn, R., & Irwing, P. (2008). Sex differences in mental arithmetic, digit span, and g defined as working memory capacity. *Intelligence 36*, 226–235.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22*, 719–748.

Meade, A. W., Lautenschlager, G. J., & Hecht, J. E. (2005). Establishing measurement equivalence and invariance in longitudinal data with item response theory. *International Journal of Testing*, *5*(3), 279–300. doi:10.1207/s15327574ijt0503_6.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics*, *7*(2), 105–118. doi:10.3102/10769986007002105.

Mendes-Barnett, S., & Ercikan, K. (2006). Examining sources of gender DIF in mathematics assessments using a confirmatory multidimensional model approach. *Applied Measurement in Education*, *19*(4), 289–304. doi:10.1207/s15324818ame1904_4.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Education*, *17*(4), 297–334. doi:10.1177/014662169301700401.

Ong, Y. M., Williams, J. S., & Lamprianou, I. (2011). Exploration of the validity of gender differences in mathematics assessment using differential bundle functioning. *International Journal of Testing*, *11*(3), 271–293. doi:10.1080/15305058.2011.555574.

Ong, Y. M., Williams, J. S., & Lamprianou, I. (2013). Exploring differential bundle functioning in mathematics by gender: The effect of hierarchical modelling. *International Journal of Research & Method in Education*, *36*(1), 82–100. doi:10.1080/1743727X.2012.675263.

Penfield, R. D. (2003). Applying the Breslow-day test of trend in odds ratio heterogeneity to the analysis of nonuniform DIF. *Alberta Journal of Educational Research*, *49*(3), 231–243.

Penfield, R. D. (2010, May). *Explaining crossing DIF in polytomous items using divergent differential step functioning effects*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Denver.

Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias. In C. R. Rao & S. Sinharray (Eds.), *Psychometrics* (Vol. 26, pp. 125–167). Amsterdam: Elsevier.

Penner, A. M. (2003). International gender x item difficulty interactions in mathematics and science achievement tests. *Journal of Educational Psychology*, *95*(3), 650–655. doi:10.1037/0022-0663.95.3.650.

Potenza, M. T., & Dorans, N. J. (1995). DIF assessment for polytomously scored items: A framework for classification and evaluation. *Applied Psychological Measurement*, *19*(1), 23–37. doi:10.1177/014662169501900104.

Qualification and Curriculum Authority. (2007a). Key stage 2: Mathematics Test A, QCA reference number 275355. Retrieved from https://www.orderline.qca.org.uk/gempdf/1847214444/1847213626.pdf.

Qualification and Curriculum Authority. (2007b). Key stage 2: Mathematics Test B, QCA reference number 275356. Retrieved from https://www.orderline.qca.org.uk/gempdf/1847214444/1847213634.pdf.

Qualification and Curriculum Authority. (2007c). Key stage 2: Mental mathematics, QCA reference number 275361. Retrieved from https://orderline.qcda.gov.uk/gempdf/1847214487/1847213642.PDF.

R Development Core Team. (2011). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistic Computing. Retrieved from http://www.R-project.org.

Raju, N. S. (1988). The area between two item characteristics curves. *Psychometrika*, *54*, 495–502. doi:10.1007/BF02294403.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Denmarks Paedagogiske Institut.

Roussos, L. A., & Stout, W. F. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement*, *20*(4), 355–371. doi:10.1177/014662169602000404.

Ryan, K. E., & Chiu, S. (2001). An examination of item context effects, DIF, and gender DIF. *Applied Measurement in Education*, *14*(1), 73–90. doi:10.1207/S15324818AME1401_06.

Shealy, R. T., & 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*(2), 159–194.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*(4), 361–370. doi:10.1111/j.1745-3984.1990.tb00754.x.

Walker, C. M., & Beretvas, S. N. (2001). An empirical investigation demonstrating the multidimensional DIF paradigm: A cognitive explanation for DIF. *Journal of Educational Measurement*, *38*(2), 147–163. doi:10.1111/j.1745-3984.2001.tb01120.x.

Walker, C. M., Zhang, B., & Surber, J. (2008). Using a multidimensional differential item functioning framework to determine if reading ability affects student performance in mathematics. *Applied Measurement in Education*, *21*(2), 162–181. doi:10.1080/08957340801926201.

Williams, J., Hadjidemetriou, C., Ryan, J., & Jones, C. (1999, September). *Investigating item and group bias in the national assessment tests for mathematics in England and Wales: Gender and language factors*. Paper presented at the British Educational Research Association Annual Conference, University of Sussex, Brighton, UK.

Zumbo, B. D. (1999). *A handbook on the theory and methods for differential item functioning: Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Ottawa, Canada: Directorate of Human Resources and Evaluation, Department of National Defense.