Eğitim ve Bilim
2014, Cilt 39, Sayı 172
Geniş Ölçekli Test Uygulamaları Özel Sayısı

Education and Science
2014, Vol. 39, No 172
Large-Scale Assessment Special Issue

Gender Differential Item Functioning in Mathematics in Four International Jurisdictions

Dört Uluslararası Bölgede Matematik Alanında Cinsiyete Göre Farklı İşleyen Maddeler

Juliette LYONS-THOMAS[1] Debra (Dallie) SANDILANDS[2] Kadriye ERCİKAN[3]
University of British Columbia

Abstract

Past research has shown a male advantage in mathematics compared to females, as well as gender differences in mathematics performance by type of item. However, most past research has focused on gender differences within one cultural group. This research examines gender differential item functioning (DIF) across four jurisdictions that took part in a large-scale international assessment: Canada, Shanghai (China), Finland, and Turkey. The findings from each jurisdiction were consistent with previous research findings: selected response formats tended to favour males, while constructed response item formats tended to favour females. Similar proportions of items favoured one gender group over the other across jurisdictions, and Finland had the greatest number of items that exhibited DIF in favour of both genders.

Keywords: Differential item functioning, gender, international assessment, mathematics performance

Öz

Yapılan çalışmalar hem erkeklerin kızlara göre matematikte daha avantajlı olduğunu, hem de madde türlerine göre matematik performansında cinsiyete göre farklılıkların olduğunu göstermiştir. Bu geçmiş çalışmaların çoğunda cinsiyet farklılıkları sadece bir kültürel grupta incelenmiştir. Bu çalışmada ise, cinsiyete göre farklı işleyen maddeler geniş ölçekli uluslararası uygulamalara katılan dört farklı bölgede incelemektedir: Kanada, Çin (Şanghay), Finlandiya ve Türkiye. Her bir bölgeye ait bulgular daha önceki araştırmaların sonuçlarını desteklemektedir: genel olarak, çoktan seçmeli soruların erkekler lehine, cevabın yazılmasını gerektiren türdeki soruların ise kızlar lehine çalıştığı görülmüştür. Bölgeler arasında, kızların lehine işleyen soru oranının erkeklerin lehine işleyen soru oranına yakın olduğu, cinsiyete göre farklı işleyen madde sayısının ise en fazla Finlandiya'da olduğu görülmüştür.

Anahtar Sözcükler: Farklı işleyen madde analizi, cinsiyet, uluslararası testler, matematik performansı

[1] Juliette Lyons-Thomas, Doctoral candidate; Dept. of ECPS, 2125 Main Mall, University of British Columbia, Vancouver, B.C. Canada V6T 1Z4; jlt319@interchange.ubc.ca

[2] Debra (Dallie) Sandilands, Doctoral candidate; Dept. of ECPS, 2125 Main Mall, University of British Columbia, Vancouver, B.C. Canada V6T 1Z4; sandilan@mail.ubc.ca

[3] Kadriye Ercikan, PhD, Professor; Dept. of ECPS, 2125 Main Mall, University of British Columbia, Vancouver, B.C. Canada V6T 1Z4; kadriye.ercikan@ubc.ca

Introduction

Researchers have examined differences in the academic performance of separate groups since the early years of testing. Often statistically significant differences are noted between groups, for example between girls and boys on a mathematics assessment (Ercikan, McCreith, & Lapointe, 2005). However, it is not always known whether those statistical differences reflect true differences in ability, knowledge and skills between the groups or whether potential problems with the test or the testing situation may bias test results, and therefore diminish the validity of drawing comparisons across groups (Ercikan & Oliveri, in press). Item bias can occur when a characteristic of the item that is not relevant to the test purpose differentially influences responses of examinee groups (Ercikan & Lyons-Thomas, 2013). There is an expectation that if an item on a test is not biased, then examinees from two groups who have equal overall ability ought to have the same probability of correctly responding to it. When examinees from different groups that have comparable ability levels have different probabilities of getting an item correct, differential item functioning (DIF) is said to occur (Hambleton, Swaminathan, & Rogers, 1991). Statistical analyses can be conducted to detect whether items on a test may be functioning differently for two groups. DIF studies typically examine policy-relevant groups based on age, gender, race or ethnicity, culture, language, country or disability. DIF signals the *potential* for bias in an item, but does not definitively establish the presence of bias. Further studies of the item are required to determine whether it is biased; such studies are often conducted through expert reviews of the item (Ercikan et al., 2010).

It is important to analyze whether items have DIF for at least two reasons. The presence of DIF signals potential bias, and bias has an impact on validity of inferences drawn from group comparisons. Therefore DIF items, if confirmed to represent underlying bias, are often removed from future administrations of a test. Second, items that exhibit DIF may have implications for curriculum and instruction (Lane, Wang, & Magone, 1996), particularly if no reason for bias can be found. For example, test items that are presented in a multiple choice item format may consistently exhibit DIF favouring one group, whereas items presented in constructed response item format may favour another group. In this case, if no bias is established, it may be desirable to ensure that all groups receive adequate instruction in completing all types of test items, and that tests contain balanced proportions of various item types. DIF should be distinguished from differences in overall group ability. For instance, the mean percent correct score for each group could be determined in order to compare groups. However, this method of comparing groups differs from DIF because it does not match on ability level.

International large-scale assessments, such as the Programme for International Student Assessment (PISA), allow researchers to investigate academic achievement and group membership from a variety of different viewpoints. For instance, in PISA not only are student achievement in reading, mathematics, and scientific literacy all assessed, but key demographic information is also gathered through questionnaires of students, their parents, and school principals. As well, sixty-five countries or economies participated in the most recent administration of PISA, which means that examinees can be compared based on a wide variety of cultural and linguistic backgrounds.

The purpose of the current research is to investigate trends and differences in gender DIF in mathematics among four jurisdictions that participated in the 2009 administration of PISA. Studies of DIF in mathematics across genders within one country or jurisdiction are frequently reported in the literature, however few studies have been conducted to compare the results of gender DIF studies in mathematics across countries or jurisdictions. PISA reports that there is a significant gender gap in mathematics across the OECD countries that took part in 2009, with 3.4% of girls and 6.6% of boys being ranked as top performers (OECD, 2010a). This paper examines to what extent DIF accounts for the gender gap in four diverse jurisdictions: Canada, Shanghai, Finland, and Turkey. In Canada and Turkey, boys performed statistically significantly higher than girls in mathematics, while in Finland and Shanghai there were no significant differences between boys' and girls' mathematics achievement (OECD, 2010b). These four jurisdictions were chosen because they are diverse in terms of culture, geographic location and proportions of socio-economically disadvantaged children. These jurisdictions were also chosen because their mathematics achievement results represent a range of scores with Shanghai, Finland and Canada ranking 1st, 4th and 9th respectively among all jurisdictions taking part in PISA 2009 and Turkey ranking 41st.

Past research points to gender differences in math items related to item type, and therefore this paper will focus on patterns of gender DIF by item type. We seek to identify differences and similarities in gender DIF across item types and across the four jurisdictions. For instance, item formats that favour one gender group over another in one jurisdiction but not another will be noted, as will those items that appear to be problematic across jurisdictions.

The following section will provide a review of the literature that investigates mathematics gender DIF within single countries as well as studies investigating gender DIF from an international perspective. This is followed by the study methodology, results, and discussion.

Review of the Literature

Gender Performance

The assertion that male examinees have an advantage on mathematics assessments has been investigated by a number of researchers on a variety of national and international tests (Liu & Wilson, 2009a). However, much of this research has been conducted on samples within a single country or jurisdiction. For instance, Liu and Wilson (2009a) used PISA 2000 and PISA 2003 math data from American examinees to examine gender DIF. More specifically, the authors examined trends in item domain and item type, finding that males had a steady advantage over female examinees. In particular, males outperformed females on complex multiple choice items (CMC) and in the space and shape domain. Lane, Wang, and Magone (1996), also studying an assessment administered in the United States, found that questions that favoured males included figures and, consistent with previous research, had to do with geometry. With regard to items that favoured females, the authors found that a feature that contributed to gender DIF was the degree to which students were asked to show their work. Lane et al. (1996) also found that females provided more conceptual explanations. The authors state that their findings are in line with other research that shows that male examinees are likely to prefer non-verbal modes of representation whereas females prefer verbal modes. Additionally, Lane et al. (1996) acknowledge that while previous research shows that male examinees do better than female examinees when answering real world scenario items, their research did not show this difference. However, the authors note that both female and male students from this study may have received classroom instruction that regularly utilized applied problems.

Working exclusively with Turkish data, Berberoglu (1995) found that one math assessment favoured males on computation items, but favoured females on both word problems and geometry items. The author also points out that because the results are from groups of students that have similar high school curricula, DIF results are not simply a result of different school programs. Examining data from Malaysian examinees, Abedalaziz (2010a) found that geometry and "real world reference" type items favoured males whereas algebraic-type questions favoured females. Another paper by the same author found that numerical ability math items favoured female examinees, whereas items involving special and deductive abilities favoured males (Abedalaziz, 2010b).

International Gender Performance

Some research has also been conducted that examined gender DIF from an international perspective. For instance, Liu and Wilson (2009b) examined PISA 2003 data and compared gender DIF between the United States and Hong Kong. The authors found that males from both countries performed better on CMC items. Males had superior performance on items that had to do with geometry and space and shape, while females showed higher performance on algebra, probability, and reproduction questions. The authors also found that while examinees from Hong Kong outperformed their American counterparts, the Hong Kong students held a wider gender gap.

Though not investigating math, Le (2009) used the 2006 PISA results to compare gender DIF among 60 test language groups. That research found that gender DIF from the international sample was highly correlated with individual language groups. That is, the author concluded that the items that showed DIF in the international sample were also more likely to have similar gender DIF within the language groups. The author also found that almost a quarter of items showed gender DIF for both genders for different language test comparisons. In other words, 24% of items could show DIF in favour of one group for one comparison of countries, and the same item could show DIF in favour of the other group for another comparison. As some of those items did not show DIF with the international sample, the author concludes that it may be necessary to examine DIF at the individual language group level.

DIF Methods

As described earlier, DIF is said to occur when different groups of examinees have different probabilities of correctly answering an item, despite comparable ability levels. There are a variety of statistical methods for identifying DIF items that vary in their approach, and sometimes, in their results (Abedalaziz, 2010b; Mendes-Barnett & Ercikan, 2006; Ercikan, Simon, & Oliveri, 2013; Gierl, Khaliq, & Boughton, 1999). For instance, one way to detect DIF items is to use a scatterplot that is composed of p-values (proportion correct statistic for each item) for two groups, with each axis representing a group. Each item is represented in this two-dimensional space as a dot. Items that are functioning equally for each group should be equal distances from each axis. If the dots on the scatterplot deviate from a 45° line, that item that the dot represents could be further investigated for bias (Sireci, Patsula, & Hambleton, 2005). A variation of this approach, called the Delta Plot Method (Angoff, 1982, 1993), plots the delta values (normal deviates based on a scale with a mean of 13 and a standard deviation of 4) for items, and allows for a distinction to be made between genuine group differences in proficiency and bias. Another DIF approach is the Mantel-Haenszel method, developed by Holland and Thayer (1988). This method relies on a chi-squared test of independence between two groups of examinees. After matching members of each group based on their test score, the probability of success is determined for each item for each group and then compared. Items are then classified into one of three effect size levels depending on their differences. Mendes-Barnett and Ercikan (2006) used Simultaneous Item Bias Test (SIBTEST, Shealy & Stout, 1993) to capture differential dimensions captured by test items and groups of test items in a British Columbia provincial assessment. In this paper, an application of the Linn-Harnisch (LH) method to item response theory (IRT) parameters will be used to identify DIF items (Ercikan & McCreith, 2002). Details of this IRT DIF detection method are described below in the Procedures section.

The above explanation of DIF methods is by no means comprehensive, but rather serves to illustrate that there are numerous approaches to identifying DIF items, rather than one all-encompassing procedure.

Method

Instrument

PISA is an international assessment that is administered every 3 years to 15-year-old students from dozens of different countries and jurisdictions including those from the OECD as well as from outside of the OECD. In 2009, the fourth round of PISA took place, with 65 countries or economies taking part. The assessment aims to measure the knowledge and skills of examinees nearing the end of their compulsory education. In particular, the exam includes three sections in the areas of mathematics, science, and reading. In each testing cycle, one of the three domains is highlighted with the majority of questions pertaining to that subject. Due to the large number of items that are required to adequately assess these broad content domains, PISA uses a complex item sampling design in which items are divided amongst 13 booklets. The booklets are randomly assigned to students in a matrix sampling procedure. Each student is administered 1 of the 13 booklets.

For this research, the mathematics portion of the PISA 2009 assessment was used. In total, there were 35 mathematics items that were administered across the 13 booklets. The items consisted of 5 item types: (1) standard multiple choice (MC) items had either 4 or 5 option responses from which students were required to select the best one; (2) CMC items presented students with several statements and for each statement they were required to select one of several possible responses; (3) closed constructed response (CCR) items required students to construct a numeric response or a single word or short phrase within very limited constraints; (4) short response (SR) items required student to generate a response with a limited range of responses for which they would get full credit; and (5) open constructed response (OCR) items involved more writing and often required explanation or justification (OECD, 2012). Table 1 provides the number of items that corresponds to each question type.

Table 1.
Number of Items by Item Type

Item Type	# of Items
Complex multiple choice	7
Multiple choice	9
Closed constructed response	3
Open constructed response	8
Short response	8
Total	35

Sample

Data from four jurisdictions were used to conduct the analysis: Canada, Finland, Shanghai, and Turkey. Each gender group from each jurisdiction consisted of approximately 2500-3000 examinees. The sample size of each group is presented in Table 2.

Table 2.
Sample Sizes for Gender Groups by Jurisdiction

Jurisdiction	Females	Males	Total
Canada	2941	2808	5749
Finland	2954	2856	5810
Shanghai	2587	2528	5115
Turkey	2551	2445	4996

The mean score for mathematics and the gender difference for each of the jurisdictions are presented in Table 3. This table shows that Canada and Turkey had statistically significant mean score differences favouring males, while Shanghai and Finland did not have statistically significant differences between gender groups (OECD, 2010b).

Table 3.
Country Mean Scores and Gender Mean Differences

Jurisdiction	Mean Score (Standard Error)	Mean Score Difference [Males – Females] (Standard Error)	Significance (p< .05)
Canada	527 (1.6)	12 (1.8)	Significant
Finland	541 (2.2)	3 (2.6)	Not Significant
Shanghai	600 (2.8)	-1 (4.0)	Not Significant
Turkey	445 (4.4)	11 (5.1)	Significant

Procedure

Each jurisdiction was analyzed separately for gender DIF using the software program Pardux (Burket, 1998). Pardux uses the LH method for identifying DIF items, which estimates parameters for the combined group as well as the focal group, and then determines fit of the combined group parameters for the focal group. A Z-score is produced that indicates the level of DIF that is present (or not present). A Z-score greater than │2.58│ with a difference between the observed and expected mean probability of getting the highest score greater than │0.1│ is identified to have level 3 DIF, or the highest degree of DIF. A Z-score that is greater than │2.58│but with a difference in observed versus expected mean less than │0.1│ is identified as level 2 DIF. Finally, if the Z-score is less than │2.58│, then the item is identified as level 1 DIF which indicates that DIF is not present. The LH method has the benefit of being able to detect DIF from data that is matrix-sampled, which as described above, is a characteristic of PISA data. In addition to determining DIF using the LH method, item characteristic curves (ICCs) were generated using the program IRT Lab (Penfield, 2003). The x-axis of an ICC represents a measure of overall ability (referred to as latent trait), and the y-axis represents the probability of correctly responding to an item. The shape and position of the curve represents the difficulty, discrimination, and guessing parameters of an item. In particular, the slope of the curve shows the degree to which the item discriminates examinees of varying ability levels. In other words, the steeper the slope of the curve, the more discrimination the item provides. A curve that is farther to the right represents a more difficult item. Finally, the point on the y-axis at which a curve begins represents the guessing parameter. A curve that is higher on the y-axis represents an item in which a respondent is more likely to guess and answer the item correctly. When the ICCs for two groups are different, DIF may be present.

Finally, two assumptions of the IRT model used in this study are that only one main trait or ability is being measured by the test (referred to as the assumption of unidimensionality) and the related assumption that conditional on ability level responses to items are independent of each other (referred to as the assumption of local independence). The failure to meet these assumptions when analyzing for DIF may lead to inaccurate results. One way to test whether these assumptions hold is to use the Q3 statistic (Yen, 1984, 1993). A Q3 statistic is calculated for every pair of items in the test. This statistic captures local item dependence and provides an indication of whether the residual error terms for the pair of items are correlated after conditioning on ability. If the local independence assumption is met, no correlation between residual terms would be expected. We used the Q3 statistic to examine the degree to which these assumptions were met in each jurisdiction included in this study.

It is also important to check whether data being analyzed fits the IRT model. One way of doing so is to examine the degree to which individual items fit the model. This can be done by comparing the scores that are actually observed for each item with the scores that would be predicted by the IRT model, using Yen's (1981) Q1 fit statistic. The Q1 is a chi-square statistic used to test the goodness of fit between the IRT model and the test data. The consequences of poor fit are that the item and ability parameters may not be estimated accurately. We used the Q1 statistic to examine the degree of fit between each jurisdiction's data and the IRT model.

Results

Tests of Assumptions and Model-Data Fit

Yen's (1984, 1993) Q3 statistic was used for assessing the assumptions of unidimensionality and local item dependence for each jurisdiction separately. Items are flagged if the statistic is greater than 0.2. Canada showed no statistically significant local item dependence, however Finland, Shanghai, and Turkey all showed significant local item dependence between items 11 and 15. The Q3 values were 0.293, 0.234, and 0.201 for Finland, Shanghai, and Turkey respectively. Unfortunately, these items have not been released by PISA, so it is not possible to determine how the two items are related. Since LID was only for one pair of the items and minimal in each jurisdiction, its effect was expected to be minimal (Yen, 1993) and these items were included in the analyses.

Fit of the data to the IRT model was also checked for each jurisdiction using the Q1 statistic (Yen, 1981). Very small proportions of items (6%) were identified as having poor fit in the Canadian, Finnish and Chinese data, with Z-scores between 5.30 and 19.28. A larger proportion of items were found to have poor fit in the Turkish data. Twelve of the 35 items (34%) had Z-scores between 4.30 and 19.22, indicating somewhat poor fit. Fit was further investigated for those 12 items by examining observed minus expected scores. Though Z-scores for the items were 4.3 or greater, and thus flagged as having poor fit, the observed minus expected values were all close to zero. Given these very small differences, the items were included in the analyses.

DIF Items

The number and percentage of DIF items for each jurisdiction are presented in Table 4. Though past research has indicated a male advantage in math assessments (Liu & Wilson, 2009a), this research found that a slightly greater number of DIF items favoured female examinees, though generally speaking, the direction of bias was relatively equivalent. The Canadian sample had eight items that showed DIF, with four favouring males and four favouring females. The Shanghai sample had five DIF items, with one more item that favoured females than males (three and two, respectively), and the Finnish sample had 11 DIF items, with one more item that favoured females over males (six and five, respectively). The Turkish sample had five items that favoured females and three items that favoured males, totalling eight DIF items.

One of the intentions of this research was to investigate the differences in gender DIF among the four jurisdictions. As Table 4 shows, the number of DIF items did vary. In particular, the Finnish sample showed 11 DIF items, while the Shanghai sample showed the least with only five DIF items. All of the items that exhibited DIF were found to be level two (moderate).

Table 4.
Number and Direction of DIF Items

Jurisdiction	Favours Females	Favours Males	Total
Canada	4 (11%)	4 (11%)	8 (23%)
Finland	6 (17%)	5 (14%)	11 (31%)
Shanghai	3 (9%)	2 (6%)	5 (14%)
Turkey	5 (14%)	3 (9%)	8 (23%)

Note. All DIF was at level 2

Another purpose of this research was to investigate if DIF varied by item type in each jurisdiction. In other words, the research sought to investigate if certain item types were more likely to be DIF for a particular jurisdiction or jurisdictions. Table 5 presents the number of DIF items for each jurisdiction by item type. Given the small pool of items, it is difficult to make a judgment about patterns, however, it is noted that SR format had the greatest number of DIF items overall.

Table 5.
Number of Differential Item Functioning Items by Item Type

Jurisdiction	Item Type					
	CMC	MC	CCR	OCR	SR	Total
Canada	2	2	0	1	3	8
Finland	1	3	1	3	3	11
Shanghai	2	0	0	1	2	5
Turkey	0	2	2	2	2	8
Total	5	7	3	7	10	

Breaking down DIF items into whether they favoured male or female examinees, Table 6 presents the number of DIF items by both type and gender. That is, whether or not items showed patterns of DIF favouring males and females according to the way in which the question was presented. This table shows that two item types exclusively favour one gender: CMC items favour only males and SR items favour only females. Additionally, MC items appear to favour male examinees to a greater extent than female examinees, with five items favouring males and only 2 items favouring females.

Table 6.
Number of Differential Item Functioning Items by Gender and Item Type

Jurisdiction	Item Type									
	CMC		MC		CCR		OCR		SR	
	F	M	F	M	F	M	F	M	F	M
Canada	0	2	1	1	0	0	0	1	3	0
Finland	0	1	1	2	1	0	1	2	3	0
Shanghai	0	2	0	0	0	0	1	0	2	0
Turkey	0	0	0	2	1	1	2	0	2	0

Note. F=Favours females; M=Favours males

Another purpose of this research was to investigate whether or not the same items were identified as DIF in all four jurisdictions. Upon review, it became apparent that a number of items functioned differentially for multiple jurisdictions. Items that were found to be DIF for at least two of the four jurisdictions are presented in Table 7. It is noteworthy that item 30, a SR type item, was found to favour female students for all four jurisdictions.

Table 7.
Items Exhibiting DIF for Multiple Jurisdictions

Item	Canada	Finland	Shanghai	Turkey	Type
1	M			M	MC
5	M	M	F		OCR
12	F	F			SR
16		F		F	CCR
17			F	F	SR
20		M		F	OCR
21	F	F			SR
23	M	M			CMC
30	F	F	F	F	SR

Note. F=Favours females; M=Favours males

Item Characteristic Curves (ICCs)

Rather than present all of the ICCs for the items that showed DIF, two ICCs are presented for the purpose of demonstration. These ICCs were selected because they demonstrate, first, an item in which there is a clear male advantage within the Canadian sample, and then second, an item in which there is a clear female advantage within the Canadian sample.

The first example is from item 7, which is a CMC item. The ICCs for Canadian male and female respondents are displayed in figure 1. This ICC shows that when males and females have similar overall ability (such as when both have a latent trait value of 1), males have a greater probability of correctly responding to the item. This trend is evident across all ability levels although it is more pronounced in the middle region of the ability scale.

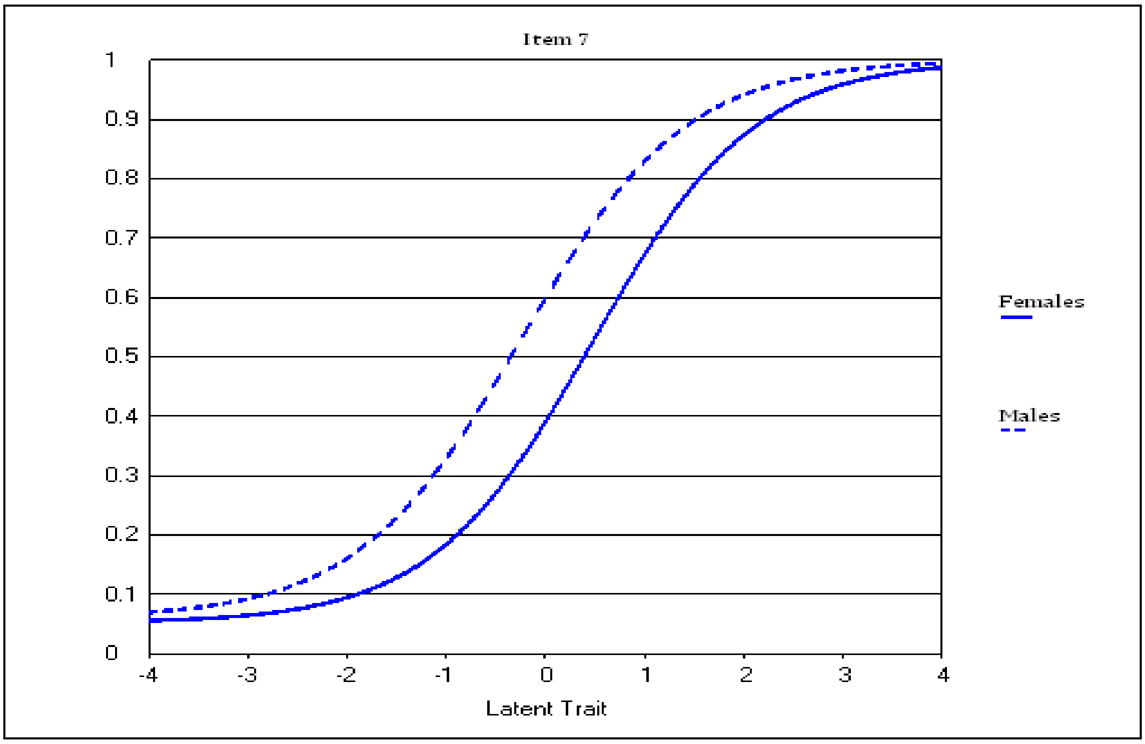

Figure 1. ICC for Item 7 displaying DIF in favour of males.

Discussion

Much of the past research points to an overall male advantage in mathematics. However, the results of this study showed that DIF items favoured both males and females, and the proportion of items favouring each group was almost equal. The findings that will be discussed shortly point to differences in item type that are consistent with previous research findings. This finding indicates that the format of math assessments may affect examinee performance on assessments differentially for gender groups.

The results of this study provide evidence that the gender gap reported by PISA for Canadian and Turkish students cannot be attributed entirely to DIF. Canada and Turkey both had a moderate number of DIF items compared to the other two jurisdictions, yet they had statistically significant gender difference in mean score, and the DIF items that were identified were relatively balanced in favouring males and females. The interpretation of differences in gender group scores requires more detailed analysis of high and low performing students. Even when there are overall group differences, the score distributions of groups overlap, which invalidates claims such as boys are outperforming girls or vice versa (Ercikan, 2009; Ercikan, Roth & Asil, in press).To contrast Figure 1 with an item that showed advantage for female examinees, Figure 2 displays the ICCs for Canadian male and female respondents for item 30, which is an SR type question. The ICC shows that females are advantaged across all ranges of the ability scale.

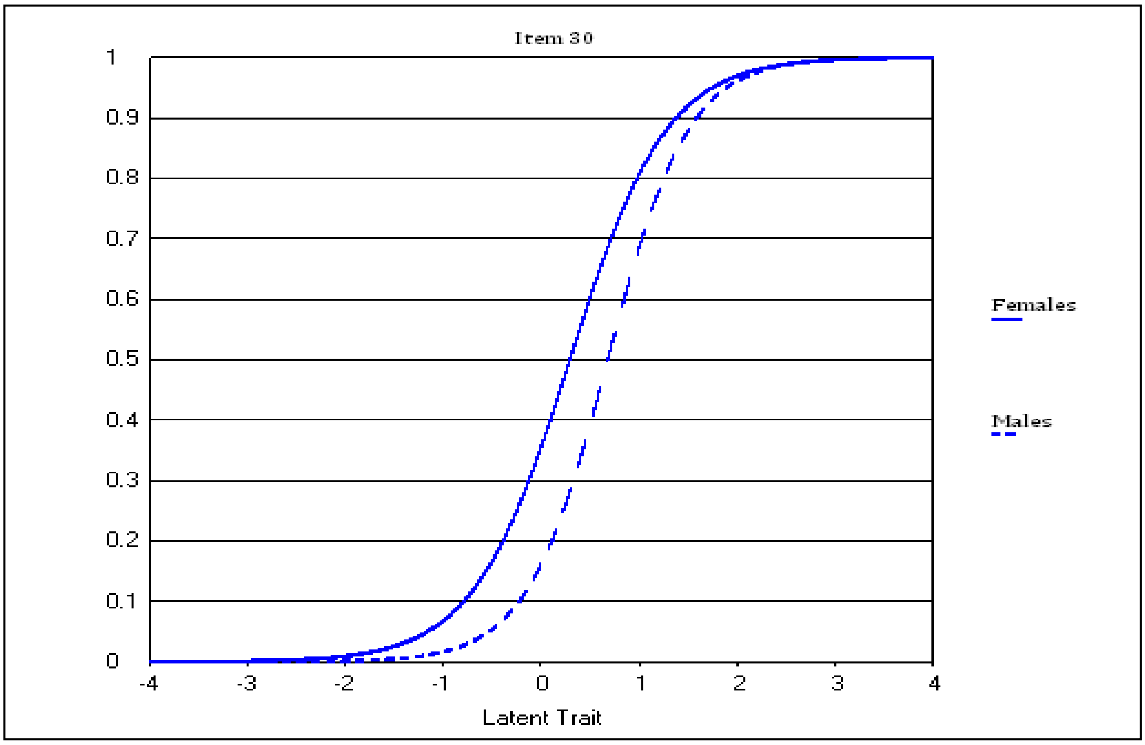

Figure 2. *ICC for Item 30 displaying DIF in favour of females.*

Another finding of the study was that Finland showed the highest number of DIF items, followed by Canada and Turkey, and Shanghai had the least number of DIF items. It is important to bear in mind that the results from this research are based on a limited number of items that were included in the mathematics portion of PISA 2009. Future research could utilize other data that includes more mathematics items, such as the 2012 administration of PISA.

A major outcome of the study was confirmation of previous research findings that there is a male advantage on MC and CMC type items, while there is a female advantage on SR type items. As described in the results section, all CMC items that showed DIF were in favour of males. Additionally, most of the MC items favoured males. Conversely, all of the SR items that showed DIF were in favour of females. Furthermore, this finding was consistent across all of the four jurisdictions. The remaining item types did not show a consistent pattern of favouring one gender group over the other; however, as mentioned above, the limited number of items included in the analysis may have contributed to this lack of consistent DIF pattern for these item types.

Identification of DIF is only the first step in investigating potential item bias (Ercikan, 2002). One future direction of this study, or any of those which examine gender DIF, would be to investigate sources of DIF. That is, while we know that some items favour one gender group over the other, we do not know the mechanism of this difference on student performance. Unfortunately, though the item type and a brief description are provided, the items themselves are not released by PISA. This type of investigation would be all the more interesting if one chose to investigate a problematic item such as item 5, which was an OCR item that favoured males in the Canadian and Finnish samples, females in the Shanghai sample, yet was not identified as a DIF item in the Turkish sample.

Another further step in this research would be to include and compare other methods of identifying DIF items, as it was mentioned above that DIF detection methods can vary in their results. It should be noted that some of the items exhibited poor fit to the IRT model that was used in this study. Therefore it may be advantageous to re-analyze the data using a non-model-based method such as the Mantel-Haenszel method to confirm the results seen here.

While some of the findings of this research are consistent with past gender DIF studies, this work also highlights the importance of examining how gender gaps vary among different nations. Though PISA is an excellent source of data for both its breadth and availability, increasing access to the types of items that students encounter could aid in the investigation of sources of DIF. These directions of research may help to lessen the observed gaps in mathematics and possibly other areas of assessment.

One major limitation to these findings is that the fit of the Turkish data to the IRT model resulted in a third of items that were flagged for poor fit. Though the observed minus expected scores were minimal, this limitation should still serve as a caution that the results from the Turkish data may not be as reliable as those from the other jurisdictions. This degree of difference in item fit points to potential differences in constructs being measured for different jurisdictions, in particular for Turkey. Further analysis could investigate test level comparability (Oliveri & Ercikan, 2011) to examine such differences.

Another limitation with respect to item fit was that items 11 and 15 were found to be locally dependant for the data sets from Finland, Shanghai, and Turkey. Though local item independence is normally required for DIF analyses, the amount of local item dependence that was found here is minimal and is therefore not expected to substantially affect the results (Yen, 1993).

Given these limitations to fit and the others previously discussed in this section, such as the limited number of items and the use of a single method to determine DIF, it is noted that the inferences from this research should be made with care. Nonetheless, the findings do point toward practical implications in the field of education. Specifically, these results can be used to inform practice. The most useful implication across jurisdictions is that there appears to be a consistent pattern in the types of items that favour gender groups. That is, males tend to outperform females on MC and CMC items, while the opposite is true for SR items. With this knowledge, teachers may choose to focus on teaching particular test strategies for different groups of examinees, and test developers may choose to ensure tests contain a balanced representation of item types. Furthermore, the findings of this study suggest that, while items that favoured one gender over another were relatively split across jurisdictions, Shanghai showed the least number of DIF items. As such, one potential avenue of inquiry could be to examine possible differences in educational instruction that may lead to Shanghai's lower gender DIF.

References

Abedalaziz, N. (2010a). A gender-related differential item functioning of mathematics test items. *The International Journal of Educational and Psychological Assessment, 5*, 101- 116.

Abedalaziz, N. (2010b). Detecting gender related DIF using logistic regression and Mantel-Haenszel approaches. *Procedia-Social and Behavioural Sciences, 7(C)*, 406-413.

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

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: Erlbaum.

Berberoglu, G. (1995). Differential Item Functioning (DIF) Analysis of Computation, Word Problem and Geometry Questions across Gender and SES Groups. *Studies in Educational Evaluation, 21*(4), 439-456.

Burkett, G. (1998). Pardux (Version 1.02) [Software]. CTB/McGraw-Hill.

Ercikan, K. (2002). Disentangling sources of differential item functioning in multi-language assessments. *International Journal of Testing*, *2*, 199-215.

Ercikan, K. (2009). Limitations in sample to population generalizing. In K. Ercikan & M-W. Roth (Eds.), *Generalizing in educational research: Beyond qualitative and quantitative polarization* (pp. 211-235), New York: Routledge.

Ercikan, K., Arim, R.,G., Law, D. M., Lacroix, S., Gagnon, F., & Domene, J. F. (2010). Application of think-aloud protocols in examining sources of differential item functioning. *Educational Measurement: Issues and Practice*, *29*, 24-35.

Ercikan, K., & Lyons-Thomas, J. (2013). Adapting tests for use in other languages and cultures. In K. Geisinger (Ed.), *APA handbook of testing and assessment in psychology, Volume 3,* (pp. 545-569). American Psychological Association: Washington, DC.

Ercikan, K., & McCreith, T. (2002). Effects of adaptations on comparability of test items and test scores. In D. Robitaille & A. Beaton (Eds.) *Secondary analysis of the TIMSS results: A synthesis of current research* (pp. 391-407). Dordrecht, the Netherlands, Kluwer Academic Publishers.

Ercikan, K., McCreith, T., & Lapointe, V. (2005). Factors associated with mathematics achievement and participation in advanced mathematics courses: An examination of gender differences from an international perspective. *School Science and Mathematics Journal*, *105*, 11-18.

Ercikan, K., & Oliveri, M. E. (in press). Is fairness research doing justice? A modest proposal for an alternative validation approach in differential item functioning (DIF) investigations. In M. Chatterji (Ed.) *Validity, fairness and testing of individuals in high stakes decision-making context* (pp. 69-86). Bingley, UK: Emerald Publishing.

Ercikan, K., Roth, W-M. & Asil, M. (in press). Cautions about uses of international assessments. *Teachers College Record*.

Ercikan, K., Simon, M., & Oliveri, M. E. (2013). Score comparability of multiple language versions of assessments within jurisdictions. In M. Simon, K. Ercikan, & M. Rousseau. (Eds.), *Improving large-scale assessment in education: Theory, issues and practice*. (pp. 110-124). New York: Routledge/Taylor & Francis.

Gierl, M., Khaliq, S.N. & Boughton, K. (1999, June). Gender Differential Item Functioning in Mathematics and Science: Prevalence and Policy Implications. Paper presented at the Canadian Society for the Study of Education.

Hambleton, R. K., Swaminathan, H. & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newberry Park, CA: Sage Publications, Inc.

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

Lane, S., Wang, N. & Magone, M. (1996). Gender-related differential item functioning on a middle-school mathematics performance assessment. *Educational Measurement: Issues and Practice, 15*(4), 21-27.

Le, L.T. (2009). Investigating Gender Differential Item Functioning Across Countries and Test Languages for PISA Science Items. *International Journal of Testing, 9*, 122–133.

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

Liu, O.L., Wilson, M. (2009b). Gender differences and similarities in PISA 2003 mathematics: A comparison between the United States and Hong Kong. *International Journal of Testing, 9*, 20-40.

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*, 289-304.

Oliveri, M., & Ercikan, K. (2011**)**. Do different approaches to examining construct comparability lead to similar conclusions? *Applied Measurement in Education, 24,* 349-366.

Organisation for Economic Co-operation and Development. (2010a). *PISA 2009 Results: Executive Summary*, Retrieved from http://www.oecd.org/dataoecd/34/60/46619703.pdf

Organisation for Economic Co-operation and Development. (2010b). *PISA 2009 Results: What Students Know and Can Do – Student Performance in Reading, Mathematics and Science, (Volume I)*. Retrieved from http://dx.doi.org/10.1787/9789264091450-en

Organisation for Economic Co-operation and Development. (2012). *PISA 2009 Technical Report,* Retrieved from http://dx.doi.org/10.1787/9789264167872-en

Penfield, R. D. (2003). IRT-Lab: Software for research and pedagogy in item response theory. *Applied Psychological Measurement, 27*(4), 301–302.

Shealy, R., & Stout, W. (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.

Sireci, S. G., Patsula, L., & Hambleton, R. K. (2005). Statistical methods for identifying flaws in the test adaptation process. In R. K. Hambleton, P. F. Merenda, & Spielberger, C. D. (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 93–115). Mahwah, N.J.: Lawrence Erlbaum Associates, Inc.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5*, 245-262.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8*(2), 125-145.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*(3), 187-213.