Investigating Gender Differential Item Functioning across Countries and Test Languages for PISA Science Items¹

Luc T. Le

Australian Council for Educational Research and
The University of Melbourne

Correspondence: Luc Le, ACER, 19 Prospect Hill Road (Private Bag 55), Camberwell VIC 3124, Australia. Email: le@acer.edu.au

¹ Paper was presented at the 5th conference of International Test Commission, Brussels, 6-8 July 2006.
Abstract

This study uses PISA cycle 3 field trial data to investigate the relationships between gender differential item functioning (DIF) across countries and test languages for science items and their formats and the four other dimensions defined in PISA framework: focus, context, competency and scientific knowledge. The data used were collected from 60 test language groups by 50 participating countries with a total of about 83,000 fifteen year-old students. An IRT method is used to detect the gender uniform DIF for each of the language groups and the whole international sample. Gender directions have been drawn for each of the five item characteristic dimensions, for example, multiple choice and closed response items tend to favour males. The study also shows the effects of the country and test language on gender DIF in the international data. Findings from this study provide a potentially valuable contribution to the development of international tests.

Key words: partial credit model, science, gender difference, international test, DIF, IRT, PISA
Introduction

PISA (Programme for International Student Achievement) study is a very large worldwide survey conducted by the Organisation for Economic Co-operation and Development (OECD). It was first conducted in 2000 and has been repeated every three years since. PISA assesses literacy in reading (in the mother tongue), mathematics and science. 57 countries participated in PISA 2006 where science was the main focus. The test had been translated (and/or adapted) into more than 40 different test languages equivalent to the English and French source versions developed by the PISA consortium. As usual, analyses of trial data (collected in 2005 for this cycle) included a procedure for detecting the presence of differential item functioning (DIF) by different demographic examinee groups for all PISA cognitive domains.

This study focussed on the analysis of gender DIF in PISA science to give comprehensive, overall pictures to identify patterns across the countries and test languages in the DIF findings, and their relationship with item characteristics defined in the PISA framework. Items identified as showing substantial DIF were not necessarily deleted from future tests, but these items were among those that needed to be carefully reviewed prior to subsequent use.

The PISA trial data has the advantage of including a large number of items, which allows greater range for exploring the relationship of the gender DIF and different dimensions of item characteristics. As shown in PISA technical reports (OECD, 2002; OECD, 2004b), PISA trial items are quite reliable for use after rigorous development and revision.
DIF Detecting Method

A variety of statistical methods for detecting DIF have been developed. Basically the procedures assume that test takers with approximately the same knowledge or ability (for example, as measured by total test scores), will perform in similar ways on individual test items across defined groups (Dorans & Holland, 1993). Popular methods are the Mantel–Haenszel statistic (Holland & Thayer, 1988), logistic regression (Swaminathan & Rogers, 1990), standardization (Dorans & Holland, 1993; Dorans & Kulick, 1986), confirmatory factor analysis and multidimensional approaches (Joreskog & Sorbom, 1989), as well as Item Response Theory (IRT; see Hambleton et al., 1991). The advantages of IRT DIF methods have been shown in many studies (see for example, Ironson & Subkoviak, 1979; Shepard et al., 1981; Shepard et al., 1985).

Two distinct forms of DIF have been recognised: uniform and non-uniform DIF. Uniform DIF is said to apply when there is no interaction between ability and group membership, or the probability of answering an item correctly is greater for one group uniformly over all ability levels. Non-uniform DIF, on the other hand, is said to apply when an interaction is found between trait level, group assignment, and item responses; or the difference in the probabilities of a correct response for the two groups is not the same at all levels of ability (Rogers & Swaminathan, 1993; Camilli & Shepard, 1994). In other words, using IRT terminology, DIF is indicated by parallel item characteristic curves and non-uniform DIF is indicated by nonparallel item characteristic curves. Detecting uniform DIF by IRT-based methods can be related to the family of Rasch model (one parameter logistic model) and its extensions (see Rasch, 1980; Wright & Masters, 1982); while IRT non-uniform DIF is related to
two or three logistic parameter IRT models and their extensions (Birnbaum, 1968; Lord, 1980; Hambleton & Swaminthan, 1985).

In the present study, only uniform DIF is detected to match the scaling IRT method applied in PISA (OECD, 2004b), where the cognitive items have been calibrated based on the Partial Credit model (Masters, 1982).

Interaction between items and country or test language

Interaction between items and countries has been mentioned in many international studies. Klieme & Burmert (2001), in a study that investigated country DIF in six countries (Austria, France, Germany, Sweden, Switzerland and the US) for TIMSS mathematics items with upper secondary students, combined with analyses of the cognitive demands of test items, showed some relative strengths and weakness of students from each of the countries by item content demands. Lapointe et al. (1992) noted that the various item formats that are used are not equally familiar to students from all countries. Using TIMSS data, O’Leary (2002) showed that the choice of item formats (multiple choice, short answer and extended response) could be one of the factors influencing the rankings of countries.

Similarly to country DIF, the source of test language DIF should be taken into account in the item selection process to ensure the equivalence of the versions of the test in multiple languages (Hambleton, 1993; Hambleton, 1994). This problem is quite common in large scale assessments. Ercikan (1999) found that 41% of science items and 18% of mathematics items from TIMSS displayed DIF when the Canadian, English and French examinees were compared. The differences may be due to problems in translation. However, other factors may affect item equivalence across language versions of tests, such as cultural and curriculum differences between the
groups (Van der Vijver & Tanzer, 1998; Ercikan et al., 2004; Gierl et al., 1999; Sireci & Berberoglu, 2000). Recent research in PISA on DIF by countries and test languages can be found in Grisay et al. (2006) and Le (2006).

**Gender difference in science**

Research on sources of DIF in science by gender has been reported in many studies. Some have focussed on item format effect (Bolger & Kellaghan, 1990; Mazzeo et al., 1993; Cole, 1997; Hamilton, 1999; Zenisky et al., 2004). Multiple choice items seem to favour males, and open-ended items tend to favour females. Some have focussed on the effect of item contents where they found that males seem to be advantaged over females on physical, earth and space science items (Becker, 1989; Jovanovic et al., 1994; Young & Fraser, 1994; Burkam et al., 1997). Males performed relatively better than females on items requiring spatial reasoning or visual content (Linn & Hyde, 1989; Halpern, 1992).

**Method**

**Data**

The data used in this study were science item-level responses by 60 test language groups from 50 participating countries (28 OECD and 22 non-OECD)\(^2\) with a total of about 83,000 students. There were approximately 49% males and 51% females. The smallest test language group contained about 450 students and the largest one contained about 5850 students.

The data included 210 science items organised in linked test forms. 91% of the items were dichotomous (scored 0 or 1) and the other 9% of the items were partial

---

\(^2\) Data from 7 other countries were not included in this study because of their late submission.
credit (scored 0, 1 or 2). Items with low discrimination or misfit to the estimation model have not been included in this analysis.

**DIF Analysis**

*Calibrating items:* Item difficulty parameter estimates by males and females from each of the studied groups are obtained separately by ACER ConQuest software (Wu et al., 1997) with the Partial Credit Model (Masters, 1982) and the parameter estimation algorithm EM (Bock & Aitken, 1981). This IRT model can handle partial credit scored items with a set of the parameters for each item containing the location of the item on the unidimensional latent continuum or item difficulty and step parameters.

*Computing DIF and Flagging DIF:* After equating the item estimates by females into the scale of item estimates by males, the DIF value for each item is computed as the difference between the two relative difficulty estimates. The corresponding chi-square test for this difference is obtained from the DIF value and the standard errors of the estimates. An item is flagged as having *substantial DIF* if the chi-square DIF test is significant at a 0.01 level and the absolute DIF value is greater than 0.25 logits. The reason for setting an additional cut point of DIF is to take into account the DIF magnitude. Any statistical test is significant if the sample size is large enough, and practically, the impact of a small DIF is not of much concern. A negative gender DIF value here means that the item is relatively in the direction of favouring males, and conversely, if the gender DIF is positive, the item is relatively in the direction of favouring females.
Item classification

Five dimensions classifying the item characteristics, which were defined in the PISA framework (OECD, 2006), were the main focus for examining the patterns of DIF in this study. The detailed category frequencies are provided in Table 1.

Insert TABLE 1 about here

Focus: Situations relating to the self, family and peer groups (Personal), to the community (Social) and to life around the world (Global).

Context: Recognising life situations involving science and technology. The classification includes: Environment (ENV), Frontiers (FRO), Hazards (HAZ), Health (HEA), Natural resources (NAT).

Competency: Explaining phenomena scientifically (EPS), Identifying scientific questions (ISQ) and Using scientific evidence (USE).

Scientific knowledge: Referring to both “knowledge of science” and “knowledge about science”. “Knowledge of science” includes Physical systems (PHYS), Living systems (LIVS), and Earth and space systems (EASS); while “Knowledge about science” refers to Scientific enquiry (SENQ), Scientific explanations (SEXP) and Science and technology (STEC).

Item format: The current PISA test consists of four types of cognitive items: (1) multiple choice (MC); (2) closed response (CR) which is short verbal or numerical response, correct answer clear-cut; (3) complex multiple choice (CMC) that is a series of true/false or yes/no choices, one answer to be chosen for each element in the series; and (4) open-ended response (OR). Most of the OR items require markers. In the IRT
analysis, the data of MC and CR items were recoded as dichotomous (0 and 1) while data from the other item types were recoded as partial credit (0, 1 and 2).

**Results**

The correlations between item difficulty estimates from male and female groups respectively at the individual test language group level vary from 0.84 to 0.95 with a mean of 0.91, and that of the whole international data is 0.96. This shows a high level of agreement between males and females on the item difficulty direction of the test.

On average of the test language groups, 5.6% of the items substantially favour males and 2.8% of the items substantially favour females. Specifically, 49% of the items are flagged as male favour in one or more test language groups, while about 35% are flagged as female favour in one or more test language groups, and 24% are shown as favouring males in at least one group and favouring females in another test language group. About 7% of the items are not flagged by any test language group.

*Item characteristics and Gender favour directions*

The percentages of substantial DIF by each of the item characteristic dimensions detected across the test language groups are presented in Table 2. Specifically, the second and sixth columns (the third and eighth columns) present the average percentage of items in each category being flagged as with male (female) favour respectively by a test language group. Columns 4 and 8 give the corresponding ratios between them (male favour versus female favour). Table 2 shows the following directions of gender favour:
Investigating Gender

Item focus: global items tend to favour males. The ratio of male/female favour in this category is more than double of that of other two categories.

Item context: ENV items seem to favour males (ratio=3.9), while NAT items tend to favour females (ratio=0.9).

Item competency: It is quite clear that EPS items would favour males (ratio=5.1), while ISQ and USE items tend to favour females (ratio=0.4 and 0.8 respectively).

Science knowledge: Males seem to be advantaged on items of “knowledge of science” (EASS and PHYS, with the ratio=12.2 and 10.4 respectively). In contrast, females seem to be advantaged on items of “knowledge about science” (SENQ and SEXP, with the ratio=0.3 and 0.7 respectively).

Item formats: It quite clear that males tend to be relatively more advantaged than females on MC/CR (ratio=5.0). The distribution of the item flags in other categories is balanced for males and females.

Agreement of international and individual test language groups on gender DIF

The (Pearson) correlations between percentages of male or female favour flagged across test language group on each item and male or female favour flagged internationally on the item (1-if flagged or 0-if not flagged) respectively are both about 0.7 ($p<<0.01$). As expected, this shows that in general the more test language group flags on male (female) favour an item gets, the more likely the item will also be internationally flagged as male (female) favour, and vice versa.
Investigating Gender

Table 3 is a crosstab table presenting the average percentage of the test language groups with male favour, female favour or none by each item where it is detected internationally as male favour, female favour or none, respectively.

*Insert TABLE 3 about here*

(a) **Consistency:** The consistency between the two approaches is shown in the figures in the shaded cells in Table 3. Each item that is flagged as male (female) favour internationally will be flagged as male (female) favour by 18.9% (12.3%) of the test language groups on average. An item that is not flagged internationally will not be flagged by 95.3% of the test language groups. Moreover, the data show that any item that is flagged internationally as female (male) favour will have at least one test language group flag in the same direction.

(b) **Opposite direction:** Each item that is flagged as male (female) favour internationally is flagged as female (male) favour by 0.4% (0.7%) of the test language groups on average. These are very small proportions compared with the corresponding consistency figures above, but indicate that the gender difference with an item for one or few test language groups may be quite different from the rest. Apart from translation problems in some cases, this may be due to cross-cultural issues.

(c) **Cancellation:** Each item that is not flagged internationally is flagged as male and female favour by 2.6% and 2.2% of the test language groups respectively. This difference may be due to one factor cancelling out the other. The item favours males in one or a few countries but favours females in other countries. An interaction between item and test language group could also be a cause.
Example: Let’s take an example of case (c), as shown in Figure 1, where an item (named as 78) is not detected as gender DIF internationally (DIF=-0.08), but is flagged as substantially gender DIF by 9 test language groups. Specifically, Figure 1 gives the item characteristic curve (ICC), which is the graph of expected scores or probability of responding correctly to the item by different ability levels, for the whole international sample, by male and female groups. The graph indicates item 78 may contain substantial non-uniform DIF at the international level, where it is likely to favour males in higher performing language groups and to favour females in lower performing language groups. This example suggests that in some cases the international DIF detecting approach may suffer a limitation when the gender DIF from test language groups varies in both directions.

Summary and discussion

Differential item functioning by sub-samples, in particular gender DIF (not equivalent to bias), is unavoidable in international tests, especially when they are used for large and heterogeneous samples, such as in PISA. Evidence here shows that gender DIF for science is dependent on item formats and content domains. The main findings in this study can be summarised as follows.

On average, for each of the test language groups, 5.6% and 2.8% of the items were flagged as substantially favouring males and females respectively. These ratios vary by the item classified categories defined in the study.

Moreover, the gender DIF detected in the international data and in the individual test language groups is highly correlated to each other. Items with
substantial DIF internationally are more likely than others to have large gender DIF in the same favour direction by the individual test language groups, and vice versa.

The consistency between the two detecting approaches is also reflected in the interaction of gender DIF with the five defined dimensions. Males tend to be relatively advantaged on MC/CR (item format); global (item focus); ENV (item context); EPS (item competency), and “knowledge of science” (science knowledge). In contrast, female favour is on ISQ (item competency) and “knowledge about science” (science knowledge). These results are similar to findings in previous studies on gender difference with item formats (see Bolger & Kellaghan, 1990; Mazzeo et al., 1993; Cole, 1997; Hamilton, 1999; Zenisky et al., 2004), and on gender difference with science item contents (see Becker, 1989; Jovanovic et al., 1994; Young & Fraser, 1994; Burkam et al., 1997).

However, the gender DIF patterns vary somewhat by the individual test language groups. Importantly, 24% of the items have been shown as both favouring males and favouring females in different test language groups. Consequently, some of these items have not been detected as gender DIF internationally. This finding suggests that examining gender DIF in individual test language groups is necessary in international tests. Items with substantial DIF by many test language groups should be of concern.

There are two potential limitations in this study: sample size effect and uniform DIF. Some DIF may have gone undetected in some individual language groups with small field-test samples. Some results from this study might be limited if inference is made to the non-uniform gender DIF, where the performance difference between males and females on an item could depend on the ability level.
References


meeting of the National Council on Measurement in Education, Montréal, Quebec, Canada.


Ironson, G. H. & Subkoviak, M. (1979). A comparison of several methods of


Investigating Gender

placement examinations (College board report no. 92-7). New York: College Entrance Examination Board.


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


Acknowledgement

The author is grateful to Professor Ronald K. Hambleton, University of Massachusetts, Amherst and Professor Jacques Gregoire, University of Louvain, Belgium for their insightful comments and thorough review of the manuscript.

Luc T. Le is Research Fellow at Australian Council for Educational Research, private Bag 55, Camberwell VIC 3124, Australia; le@acer.edu.au. His research interests include IRT and DIF methods in large-scale assessment, as well as mathematical optimisation and programming.
Footnotes

1 Paper was presented at the 5th conference of International Test Commission, Brussels, 6-8 July 2006.

2 Data from 7 other countries were not included in this study because of their late submission.
<table>
<thead>
<tr>
<th>Dimensions</th>
<th>Number</th>
<th>Per cent</th>
<th>Dimension</th>
<th>Number</th>
<th>Per cent</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Focus</strong></td>
<td></td>
<td></td>
<td><strong>Knowledge</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Global</td>
<td>49</td>
<td>23.3</td>
<td>EASS</td>
<td>24</td>
<td>11.4</td>
</tr>
<tr>
<td>Personal</td>
<td>62</td>
<td>29.5</td>
<td>LIVS</td>
<td>46</td>
<td>21.9</td>
</tr>
<tr>
<td>Social</td>
<td>99</td>
<td>47.1</td>
<td>PHYS</td>
<td>45</td>
<td>21.4</td>
</tr>
<tr>
<td><strong>Context</strong></td>
<td></td>
<td></td>
<td>SENQ</td>
<td>42</td>
<td>20.0</td>
</tr>
<tr>
<td>ENV</td>
<td>33</td>
<td>15.7</td>
<td>SEXP</td>
<td>42</td>
<td>20.0</td>
</tr>
<tr>
<td>FRO</td>
<td>71</td>
<td>33.8</td>
<td>STEC</td>
<td>11</td>
<td>5.2</td>
</tr>
<tr>
<td>HAZ</td>
<td>25</td>
<td>11.9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HEA</td>
<td>61</td>
<td>29.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NAT</td>
<td>16</td>
<td>7.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>4</td>
<td>1.9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Competency</strong></td>
<td></td>
<td></td>
<td><strong>Format</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EPS</td>
<td>104</td>
<td>49.5</td>
<td>CMC</td>
<td>55</td>
<td>26.2</td>
</tr>
<tr>
<td>ISQ</td>
<td>50</td>
<td>23.8</td>
<td>CR</td>
<td>7</td>
<td>3.3</td>
</tr>
<tr>
<td>USE</td>
<td>56</td>
<td>26.7</td>
<td>MC</td>
<td>75</td>
<td>35.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>OR</td>
<td>73</td>
<td>34.8</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>210</td>
<td>100</td>
<td><strong>Total</strong></td>
<td>210</td>
<td>100</td>
</tr>
</tbody>
</table>
### TABLE 2
A summary of percentage of gender DIF flags

<table>
<thead>
<tr>
<th>Dimensions</th>
<th>Male favour</th>
<th>Female favour</th>
<th>Ratio</th>
<th>Dimensions</th>
<th>Male favour</th>
<th>Female favour</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Focus</strong></td>
<td></td>
<td></td>
<td></td>
<td><strong>Knowledge</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Global</td>
<td>8.0</td>
<td>1.8</td>
<td>4.4</td>
<td>EASS</td>
<td>12.2</td>
<td>1.0</td>
<td>12.2</td>
</tr>
<tr>
<td>Personal</td>
<td>5.5</td>
<td>3.6</td>
<td>1.5</td>
<td>LIVS</td>
<td>5.5</td>
<td>2.7</td>
<td>2.0</td>
</tr>
<tr>
<td>Social</td>
<td>4.6</td>
<td>2.8</td>
<td>1.6</td>
<td>PHYS</td>
<td>9.4</td>
<td>0.9</td>
<td>10.4</td>
</tr>
<tr>
<td><strong>Context</strong></td>
<td></td>
<td></td>
<td></td>
<td><strong>SENQ</strong></td>
<td>1.5</td>
<td>4.7</td>
<td>0.3</td>
</tr>
<tr>
<td>ENV</td>
<td>8.5</td>
<td>2.2</td>
<td>3.9</td>
<td>SEXP</td>
<td>3.1</td>
<td>4.7</td>
<td>0.7</td>
</tr>
<tr>
<td>FRO</td>
<td>7.1</td>
<td>2.8</td>
<td>2.5</td>
<td>STEC</td>
<td>1.8</td>
<td>1.1</td>
<td>1.6</td>
</tr>
<tr>
<td>HAZ</td>
<td>5.5</td>
<td>2.6</td>
<td>2.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HEA</td>
<td>3.4</td>
<td>3.6</td>
<td>0.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NAT</td>
<td>3.4</td>
<td>2.1</td>
<td>1.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Competency</strong></td>
<td></td>
<td></td>
<td></td>
<td><strong>Format</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EPS</td>
<td>9.2</td>
<td>1.8</td>
<td>5.1</td>
<td>CMC</td>
<td>3.5</td>
<td>3.1</td>
<td>1.1</td>
</tr>
<tr>
<td>ISQ</td>
<td>1.5</td>
<td>4.1</td>
<td>0.4</td>
<td>MC/CR</td>
<td>8.0</td>
<td>1.6</td>
<td>5.0</td>
</tr>
<tr>
<td>USE</td>
<td>2.8</td>
<td>3.6</td>
<td>0.8</td>
<td>OR</td>
<td>4.7</td>
<td>4.0</td>
<td>1.2</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5.6</td>
<td>2.8</td>
<td>2.0</td>
</tr>
</tbody>
</table>
TABLE 3

Agreement between international and test language groups on DIF flags

<table>
<thead>
<tr>
<th>Average percentage by category</th>
<th>International DIF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Male favour</td>
</tr>
<tr>
<td><strong>Test</strong></td>
<td></td>
</tr>
<tr>
<td>Male favour</td>
<td>18.9</td>
</tr>
<tr>
<td>Female favour</td>
<td>0.4</td>
</tr>
<tr>
<td>Not flagged</td>
<td>80.7</td>
</tr>
<tr>
<td><strong>DIF</strong></td>
<td>Total (%)</td>
</tr>
</tbody>
</table>
Figure Captions

Figure 1. An example of international gender ICCs
Figure 1. An example of international gender ICCs