Analysis of Differential Item Functioning

Luc Le
Australian Council for Educational Research
Melbourne/Australia
le@acer.edu.au

Paper prepared for the Annual Meetings of the American Educational Research Association in San Francisco, 7-11 April 2006.

Abstract

For the development of comparable tests in international studies it is essential to examine Differential Item Functioning (DIF) by different demographic groups, in particular cultural and language groups. For the selection of test items it is important to analyse the extent to which items function differently across the sub-groups of students.

In this paper the procedures used in the 2006 field trial for investigating DIF for Science items are described and discussed. The demographic variables used are country of test, test language and gender (at a country level). Item Response Theory (IRT) is used to analyse DIF in test items. The outcomes of DIF analysis examined and discussed with reference to the item characteristics defined in PISA framework: format, focus, context, competency, science knowledge, and scoring points. The DIF outcomes are also discussed with item feedback ratings provided by national research centres of the participating countries, where items are rated according to the countries' priorities, preferences and judgements about cultural appropriateness.

INTRODUCTION

The PISA study (Programme for International Student Achievement) is a very large survey around the world conducted by the Organisation for Economic Co-operation and Development (OECD). It was first conducted in 2000 and has been repeated every three years. PISA assesses literacy in reading (in the mother tongue), mathematics and science. In 2000 reading was the major assessment domain, while in 2003 the major domain was mathematics and in 2006 the major domain will be science. In these areas PISA mainly assesses 15-year-old students' capacities to use their knowledge and skills in order to meet real-life challenges, rather than merely looking at how well they have mastered a specific school curriculum. Students have to understand key concepts, master certain processes and apply knowledge and skills in different situations. In 2000, 28 OECD Member countries and four other countries carried out the first PISA survey. A further 13 countries conducted the same survey in 2002. In 2003 there were 42 countries participating in PISA. It is expected that about 60 countries will participate in PISA 2006 where science will be the main focus. The test will be translated (or adapted) into about 40 different test languages equivalent to the English and French source versions developed by the PISA consortium.
As in usual routines, analyses of trial data for 2006, including a procedure for detecting the presence of differential item functioning (DIF) for science, were implemented. Items with a large DIF found during field testing have often been returned to item writers for review and revision or simply removed.

The goal of this paper is not to look narrowly at individual DIF items where the information has already been provided in the item selection and development process for PISA main study, but rather to give comprehensive overall pictures to identify patterns in the DIF findings.

Identifying the causes of DIF is also important part to understand about the relative strength and weakness of the examinee groups on the different skills and abilities that the test items measure. Some possible sources for such trends may include item content, item type or format, item context, content and cognitive dimensions associated with items. It may be possible to gain considerable insight into the potential causes of DIF by considering the statistical evidence of item-level DIF in light of such item attributes. Practically, items identified as showing substantial DIF are not necessarily deleted from future tests, but these items are among those that need to be carefully reviewed prior to any subsequent use.

Commonly, DIF studies have examined cognitive tests for the presence of item DIF or potential test bias with respect to a number of different demographic characteristics, such as gender, education, social class, ethnicity, age and so on. In particular, as the main purpose of selecting items for PISA 2006 main cycle, this study is concentrated on DIF analysis of science cognitive data with respect to three main variables: country, gender and test language.

In a study that investigated country DIF in six countries (Austria, France, Germany, Sweden, Switzland and the US) for TIMSS mathematics items with upper secondary students, combined with analyses of the cognitive demands of

test items, Klieme & Burmert (2001) show 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 1995 data, O'leary (1995) shows that the choice of item formats (multiple-choice, short-answer and extended response) could be one of the factors influencing the country rankings. Similarly to country DIF, study of test language DIF is very important in the item selection process. It helps to detect items that behave differently by different test language groups in order to insure equivalence of the versions of the test in multiple languages (Hambleton, 1993; Hambleton, 1994). 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,1998; Ercikan, 2002 ; Ercikan et al., 2004; Gierl et al., 1999; Sireci & Berberoglu, 2000). The DIF study can also provide useful information to understand some possible sources of those variations. Item differential functioning is often large on adapted tests (Ercikan et al., 2004).

Research on sources of DIF in science by gender has been reported in many studies. Some of them focus on item format effect (Bolger & Kellaghan, 1990; Mazzeo et al., 1993; Cole, 1997; Hamilton &Snow, 1998; Hamilton, 1999; Zenisky et al., 2004). Multiple choice items seem to favour male examinees and open-ended items tend to favour female examinees. Some focus on the effect of item content where they found that males seem to be advantaged over females on physical science items and earth and space science items (Becker, 1989; Jovanovic et al., 1994; Young & Fraser, 1994; Burkam et al., 1997). And on the effect of item cognitive domains, some evidence was found that male examinees performed differentially better than female examinees (when matched on total test score) on items requiring spatial reasoning or visual content (Linn & Hyde, 1989; Halpern, 1992).

However, in international tests, by different cultures or educational systems, the gender DIF patterns and the above effects on gender DIF varied somewhat by different examinee countries. In other words, the interaction between items or domains with gender would be different from country to country (see TIMSS 2003 report, Mullis et al., 2004).

A variety of statistical methods for detecting DIF have been developed, where basically, the procedures assume that if test takers have approximately the same knowledge or ability (for example, as measured by total test scores), then they should perform in similar (though not identical) ways on individual test items across the defined groups (Dorans & Holland,1993). Among them the popular ones are Mantel–Haenszel statistic (Holland & Thayer, 1988), logistic regression (Swaminathan & Rogers, 1990), standardization (Dorans & Holland, 1993; Dorans & Kulick, 1986), SIBTEST (Shealy & Stout, 1993), and Item Response Theory (IRT; see Hambleton, Swaminathan, & Rogers, 1991). Specifically, some IRT DIF methods are based on comparing the estimation values of item parameters or comparing the goodness of fit between item response models and the data (Thissen et al., 1993; Lord, 1980); some develop statistical tests for testing significance or measuring the difference between the curves obtained

from the two study groups (Thissen et al., 1993; Raju, 1988, 1990), such as estimating the area between the curves, or the squared differences measure, or weighting the area and squared differences measures. IRT methods have been shown in many studies to be superior to other methods (Ironson, 1977; Ironson & Subkoviak, 1979; Merz & Grossen, 1979; Runder, Getson, & Knight, 1980; Shepard, Camilli & Averill, 1981; Shepard, Camilli & Williams, 1985; Subkoviak, Mack, Ironson, & Craig, 1984). When IRT models properly describe the data, they also provide sensitive tests for DIF. However, IRT-based approaches seem unsuitable with very small sample sizes. Large numbers of individuals are needed to fit IRT curves. It was estimated by Embretson and Riese (2000) that between 250 and 500 individuals are needed for stable IRT item parameter estimates. The condition that DIF detection would require at least 250–500 individuals in all groups analysed, is fortunately satisfied in this study.

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 level 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 the IRT-based methods can be related to family of the item response Rasch models (or one parameter logistic model and its extensions; see Rasch, 1980; Wright & Stone, 1979; Wright & Masters, 1982); while detecting non-uniform DIF by IRT-based methods is related to two or three logistic parameter IRT models and their extensions (Birnbaum, 1968; Hambleton & Swaminthan, 1985).

There have been long debates about the choice of uniform DIF or non-uniform DIF in student response data from large-scale assessments (Camilli & Shepard, 1994; Zumbo, 1999). In the present study, uniform DIF is used to match the scaling IRT method applied in PISA (see PISA 2003 technical report, Adams, 2004). In PISA the item response model used for cognitive data is Partial Credit (Masters, 1982) which is an extension of the simple logistic Rasch model for tests containing polytomous items.

Objectives of the study:
Apart from providing detailed information for individual items in the selection process, the main objectives of this study are:

- Using an IRT approach to investigate uniform DIF of the science items or variations of item difficulty parameters across the examinee groups by country, gender or test language.
- Examining the DIF patterns associated with item characteristics and item feedback.

METHOD

Data

The data used in this study were science item-level responses of approximately 84,000 students collected from 50 participating countries (28 OECD and 22 non-OECD). There were about 49% males and 51% females. The detailed distribution is given in Table A1 (Appendix 1). Data from a few countries who that submitted late are not included in this study. The data include 210 science items from 14 main test booklets and 10 half-booklets (for students who participated in both paper and computer-based tests). Items with bad statistics from a primary analysis, such as low discrimination or misfit to the estimation model, have not been included in this analysis. In this trial, each main booklet contained about 50 to 60 cognitive items and the items were organised in a linked design.

DIF Analysis

The DIF analysis process is implemented in the following main steps:

Calibrating item: Item difficulty parameter estimates by each of the studied groups are obtained by Conquest (Wu et al., 1997) with the Partial Credit Model (Masters, 1982) and the parameter estimation algorithm *EM* (Bock & Aitken, 1981). The group data sets are: individual countries, each test language, separated males and females in a country level, and the whole international data set.

The Partial Credit Model can be described by a mathematical probability function:

$$P_{ix}(\theta) = \frac{\exp \sum_{j=0}^{x} (\theta - \delta_i - \tau_{ij})}{\sum_{k=0}^{m_i} \exp \sum_{j=0}^{k} (\theta - \delta_i - \tau_{ij})}, \; x = 0, 1, 2, ..., m_i$$

where $P_{ix}(\theta)$ denotes the probability of a person with ability level θ (on the latent dimension) to score x on item i with $m_i + 1$ ordered categories 0, 1,…, m_i. Here, τ_{ij} denotes a step parameter, standing for the event that the person responded to category *j rather than j-1* ($\tau_{i0} \equiv 0$). The item parameter δ_i gives the location of the item on the latent continuum. This parameter is also known as "item difficulty".

Equating scale: to make item difficulty estimates comparable for each pair of studied groups, the estimates from the "focal" group are transformed onto the "reference" group scale. Roughly speaking, let $b_1, b_2, ..., b_k$ and $a_1, a_2, ..., a_k$ be item difficulty estimates by these groups, respectively (here k=210). The first step is to compute the difference $d_i = a_i - b_i$ for each item. Then second step is to compute the mean of that difference M, which is called the "shift" value. And the

last step is adjusting the estimate values b_i of the focal group: $\overline{b_i} = b_i + M$. In order to avoid very large DIF items influencing the scale transforming, in the second step, d_l that is very different from the cohort (d_i) will not be used to compute the shift M.

In this study design:

- Country DIF: the whole international group is the reference group and individual countries are focal ones.
- Language DIF: English group is the reference group and other language groups are focal ones.
- DIF gender: in each of the countries the male group is the reference group and the female group is the focal one.

Computing DIF and DIF test: After equating the scale of item estimates from this group to the other group, the DIF value for item *i* is computed as the difference between its relative difficulty estimates by the compared groups: $DIF_i = a_i - \overline{b_i}$; and the corresponding chi-square test for this difference is obtained from the DIF value and the corresponding standard errors of the estimates.

Flagging DIF: for country and language DIF, item is flagged if the chi-square DIF test is significant at a 0.01 level and its absolute DIF value is greater than 0.30 logit; while with gender DIF, item is flagged if the chi-square DIF test is significant at a 0.01 level and its absolute DIF value is greater than 0.25 logit. The reason for setting an additional cut point of DIF is to take into account the DIF magnitude. Any statistic test would be significant if the sample size is large enough. And additionally, due to heterogenous sample problems, it is expected that in an international test, item parameter estimates would vary across countries or test languages (each may combine some countries) more than between male and female groups within a country. For example, 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.

Examining DIF patterns: The country, test languages and gender pattern are examined together with item classifications and characteristics, and with their feedback ratings from participated countries.

Framework for Item classification

Six variables or dimensions classifying the item characteristics, which were defined in the PISA framework, will be the main focus for examining the patterns of DIF in this study. The detailed category frequencies are provided in Table 1.

Focus: The focus of the items will be on situations relating to the self, family and peer groups (*Personal*), to the community (*Social*) and to life across 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: Multiple choice (MC); closed constructed-response (CR) which is short verbal or numerical response, correct answer clear-cut; 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 open constructed-response (OR). Most of the OR items require markers. Then in IRT analysis, 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. Because there are only 7 CR items (3.3%), interpretations related to them are limited.

Item maximal score point: There are 192 items with score points 0 and 1 (1-score point maximum); and 18 items with score points 0, 1 and 2 (2-score point maximum) but most are OR items. Therefore, in this dimension it is designed to work with OR items only.

Table 1. Item Classification

	Number	Percent		Number	Percent
Item Focus			**Item Competency**		
Global	49	23.3	EPS	104	49.5
Personal	62	29.5	ISQ	50	23.8
Social	99	47.1	USE	56	26.7
Total	210	100	*Total*	210	100
Item Context			**Science Knowledge**		
ENV	33	15.7	EASS	24	11.4
FRO	71	33.8	LIVS	46	21.9
HAZ	25	11.9	PHYS	45	21.4
HEA	61	29.0	SENQ	42	20.0
NAT	16	7.6	SEXP	42	20.0
Other	4	1.9	STEC	11	5.2
Total	210	100	*Total*	210	100
Item Format			**Item maximal score**		
CMC	55	26.2	1	192	91.4
CR	7	3.3	2	18	8.6
MC	75	35.7	1 - OR	59	
OR	73	34.8	2 - OR	14	
Total	210	100	*Total*	210	100

Framework for item feedback ratings

Participating countries were asked to use a five point scale (1 to 5) for rating each of the trial cognitive items on Curriculum, Relevance, Interest, Authenticity, Culture, and Priority:

Curriculum: Use rating 1 (not in curriculum) through to rating 5 (standard curriculum material) to indicate how close the item is to school curriculum. In the case where school curriculum varies within the country, base the ratings on the percentage of students who would have covered the content area of the item.

Relevance: How relevant is the task for the students in preparing for life? It refers to skills that are needed in many facets of life, such as for work, pleasure, and participation in society. Use rating 1 for 'not relevant', through to rating 5 for 'highly relevant'.

Interest: How interesting is the task for the students? Here it refers to motivational aspects of the task. Use rating 1 for 'not interesting', through to rating 5 for 'extremely interesting'.

Authentic: Will students regard the context as authentic? Does it involve an authentic application of science and/or technology? This does not mean that the students would necessarily encounter the context in their everyday life.

Cultural (*concern*): Are there any cultural concerns about the items? Answer Yes (1) or No (0). This question refers to problem settings that are not acceptable in the country, for reasons such as religion or culture. For example, the promotion of the legalisation of drug use may not be acceptable as the context for an item.

Priority (for inclusion): Use '1' for showing a low priority for that item's inclusion in the main survey, '2' for showing moderately low priority, '3' medium level priority, '4' moderately high priority, through to '5', for showing that the unit/item is of highest priority for inclusion.

For each of the above categories, the rating index used in this study is the mean of the country ratings.

RESULTS

The percentages of DIF flags in total and by each of the item characteristics are presented in Table 2. In general, this table shows that the gender DIF with an average of 10% is lower than the country DIF (25%) and the language DIF (39% to 59%). The item performances by the English group are closest to those by the French group (39% DIF flags); then by the Spanish group (44% DIF flags); and followed by the German, Russian and Arabic groups with 48%, 53% and 59% DIF flags, respectively. Moreover, interestingly, although the country DIF flags for OCED members are lower than those for all PISA members (22% compared

with 25% % on average), the gender DIF flags for OCED members are slightly higher than those for all PISA members (12% compared with 10% in average).

Table 2 Average DIF Flags

	Percent country DIF Flag	Percent OECD country DIF Flag	Percent gender DIF Flag	Percent OECD gender DIF Flag	Percent Arabic DIF Flag	Percent French DIF Flag	Percent German DIF Flag	Percent Russian DIF Flag	Percent Spanish DIF Flag
Focus									
Global	24	22	12	13	53	41	55	49	53
Personal	26	24	11	13	65	34	32	53	37
Social	24	21	9	10	59	41	54	55	43
Context									
ENV	27	24	13	15	64	45	48	58	45
FRO	23	20	11	14	61	44	46	46	42
HAZ	28	28	10	11	52	28	40	44	44
HEA	24	22	8	10	61	36	49	57	44
NAT	25	23	6	8	56	31	50	63	50
Competency									
EPS	26	23	13	15	56	37	49	55	51
ISQ	24	22	7	8	76	42	48	62	42
USE	23	22	7	9	50	41	45	41	32
Science Knowledge									
EASS	26	24	15	17	71	42	38	42	42
LIVS	26	24	10	12	59	37	57	54	52
PHYS	25	22	12	14	47	33	49	53	47
SENQ	23	20	8	9	79	43	40	62	45
SEXP	21	19	9	11	48	36	43	40	36
STEC	31	31	3	4	55	64	73	82	27
Format									
CMC	23	21	8	9	65	35	49	53	33
CR	22	17	25	31	57	57	29	29	43
MC	22	20	10	12	52	43	51	60	44
OR	29	26	10	12	62	37	45	48	52
Maximal score									
1	24	22	9	11	61	39	49	54	45
2	28	26	16	19	39	44	33	39	28
1 – **OR**	29	26	9	11	68	37	49	51	59
2 – **OR**	28	28	14	16	36	36	29	36	21
Overall	25	22	10	12	59	39	48	53	44

1. Item Differential Functioning by Country

The main purpose of investigating country DIF here is to see if there is any relationship between the variations of item difficulty across countries and item characteristics.

The correlation of the item difficulty estimates by individual country with those by the international data varies from 0.72 to 0.96 with a mean 0.88, where the lowest indices belong to some new PISA country members: Azerbaijan, Tunisia,

Jordan and Taipei, and the highest ones belong to OECD members. This figure shows that in general the orders of item difficulties by country are quite consistent.

The variation of item difficulty estimates across countries for each item is computed by the mean of the absolute values of the individual country DIF (i.e. the country estimated value against the corresponding international value), and is named as (unsigned) *country DIF* for the item. The distribution of the *country DIF* for the 210 items has a mean of 0.36 logits and standard deviation of 0.10.

a) Country DIF and item difficulty

The correlation between international item difficulty estimates and the country DIF is $r = 0.22$ (p=0.001), which is significant at the 0.01 level. This small correlation suggests a slight tendency for more variation in item difficulty across countries for harder items than easier ones.

b) Country DIF and item characteristics

Additionally to Table 2, Figure 1 provides box-lot graphs of the (unsigned) country DIF by each of the item characteristic variables.

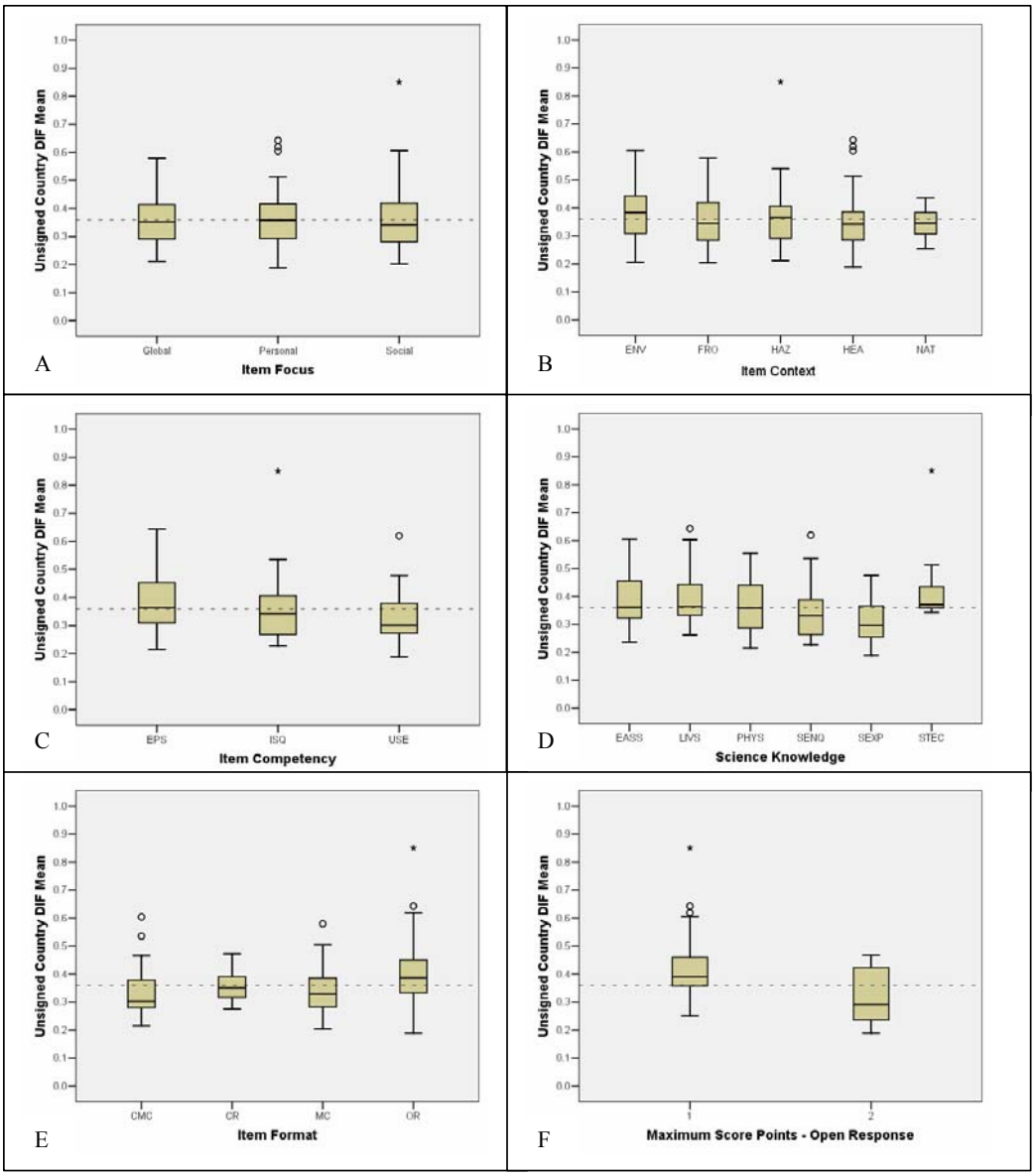
- *Focus*: The country DIF seems to be similar in the 3 categories (Figure 1A).

- *Context*: Items in NAT, HEA and FRO tend to be more stable than items in ENV and HAZ (Figure 1B).

- *Competency*: USE items are most stable, and next are ISQ ones. EPS items tend to have more country DIF (Figure 1C).

- *Scientific knowledge*: Figure 1D indicates that "knowledge of science" items tend to have more country DIF than "knowledge about science" items, except STEC. Among "Knowledge of science" items, PHYS seems to have slightly lower country DIF problems than the other two categories. Among "Knowledge about science" SEXP items appear to be most stable, while STEC items are least stable.

- *Item format*: Figure 1E shows more country DIF in OR items than other items. A large proportion of the OR items require markers, could be a factor in the variation across countries. Among OR items, the item difficulty is more likely to be stable across countries for 2-score point items than 1-score point items (Figure 1F).



(Circles - outliers, stars - extreme cases, dash line – group median, and dotted line - overall mean reference)

Figure 1. Country DIF and Item Characteristics

c) Country DIF and country feedback ratings

There is no significant correlation at a 0.05 level (the correlation is very close to 0) between the country DIF and the country mean of the rating variables: Curriculum, Relevance, Interest, Authenticity, Culture, and Priority.

2. Item Differential Functioning by Test Language

The item difficulties estimated from English groups are compared with those for French (FRE), Spanish (SPA), German (GER), Arabic (ARA) and Russian (RUS). These were the largest test language groups in the PISA trial 2006 data. The distribution of these language groups is given in Table A2 (Appendix 1). English and French are the source versions developed by the PISA consortium. The correlations of the item difficulty estimates between these groups are presented in Table 3. All of these correlations are significant at 0.01 level (p=0.000). The table shows that in general, the consistencies of item difficulty trends for English, French, German and Spanish groups are very high, with correlations of about 0.90. In contrast, Arabic and then Russian groups seem to be a bit different from those groups. The correlations of item difficulty estimates by the Arabic group and by other groups vary from 0.76 to 0.83, and the correlations of item difficulty estimates by the Russian group and by other groups vary from 0.80 to 0.87.

Table 3. The correlations of item difficulty estimates by language

	English	Arabic	French	German	Russian	Spanish
English	1					
Arabic	0.80	1				
French	0.91	.79	1			
German	0.90	.76	.89	1		
Russian	0.81	.80	.84	.85	1	
Spanish	0.90	.83	.91	.89	.87	1

a) Test language DIF and item difficulty

The correlations between the item difficulty estimates by English group and the corresponding absolute value of language DIF (against English) for Arabic, French, German, Russian and Spanish are 0.07 (p=0.30), 0.21 (p<<0.01), 0.03 (p=0.67), 0.10 (p=0.14) and 0.10 (p=0.14), respectively. This does not provide a relationship between item difficulty and language DIF yet.

b) Test language DIF and item characteristics

Figure 2 presents line graphs of the DIF between English and the other five languages DIF, relating to item characteristics. In the DIF computation here, a negative value means that the item favours the English group, while a positive value means that the other language group is favoured. Roughly speaking, the graph shows that Arabic is most different from English, and second is Russian, while the DIF between French and English is smallest.

- *Focus:* There is a small difference in the DIF means by each category. The English group tends to find slightly more difficulty with "Global" items than the other language groups. The Arabic and Russian groups seem to find "Personal" items relatively harder.

- *Context:* Among five stated categories, the DIF means for the five languages are close to zero at ENV and HEA. German and Spanish tend to find NAT slightly easier. The Arabic group finds FRO relatively harder, but HAZ relatively easier.

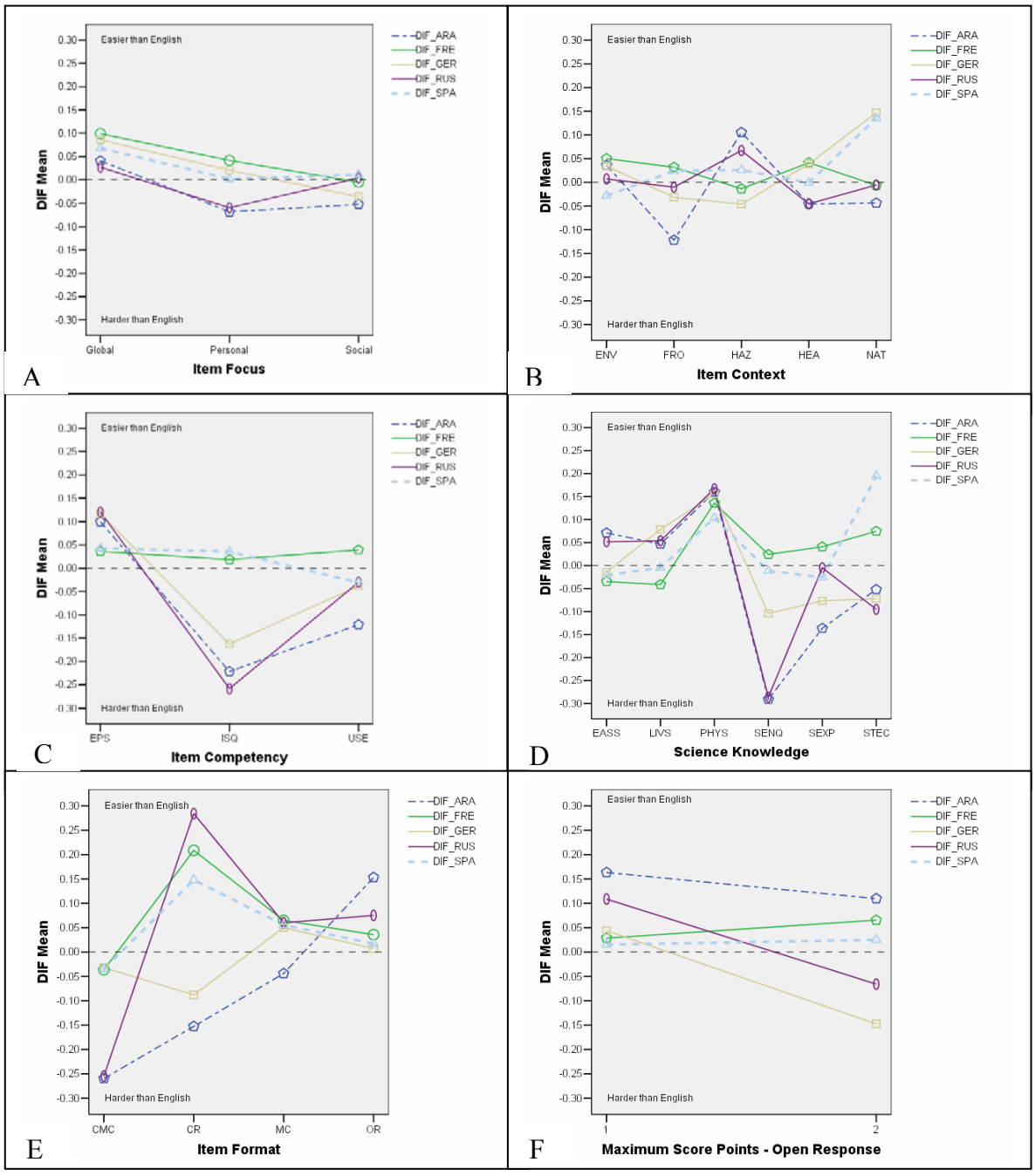


Figure 2. Test Language DIF and Item Characteristics

- *Competency*: The means of French and Spanish DIF on three categories are close to zero, while Arabic, German and Russian DIF patterns are in the same direction. ISQ items and then USE items are relatively harder for them; while EPS items are easier for them than for the English group.

- *Scientific knowledge:* The five language groups are close to the English group in EASS and LIV items, but they tend to do relatively better than English in PHYS items. Arabic, Russian and German groups are relatively less able in "knowledge about science", especially with SENQ, but they seem to be more favoured with "Knowledge of science". The Spanish group, on the other hand, tends to do better on STEC items.

- *Item Format:* French, Spanish and Russian score relatively highly in CR. While CMC items seem to be relatively easier in English than in the other five languages, OR seems to be slightly harder in English than in the others. Moreover, items in CMC format seem relatively harder in Arabic and Russian. The Arabic group finds CR slightly relatively harder as well. Within OR items, items with 2-score points seem to be relatively more difficult for the German and Russian groups than for the English group.

c) Test language DIF and country feedback ratings

There is no significant correlation found at the 0.05 level (the correlations tend to be close to 0) between the five language DIFs and the rating variables: Curriculum, Relevance, Interest, Authenticity, Culture, and Priority.

3. Item Differential Functioning by Gender

The difference of item difficulty estimates by males and females for each country is investigated by gender DIF. In this study, a negative gender DIF means that the item tends to favour males and conversely, if the gender DIF is positive, the item tends to favour females.

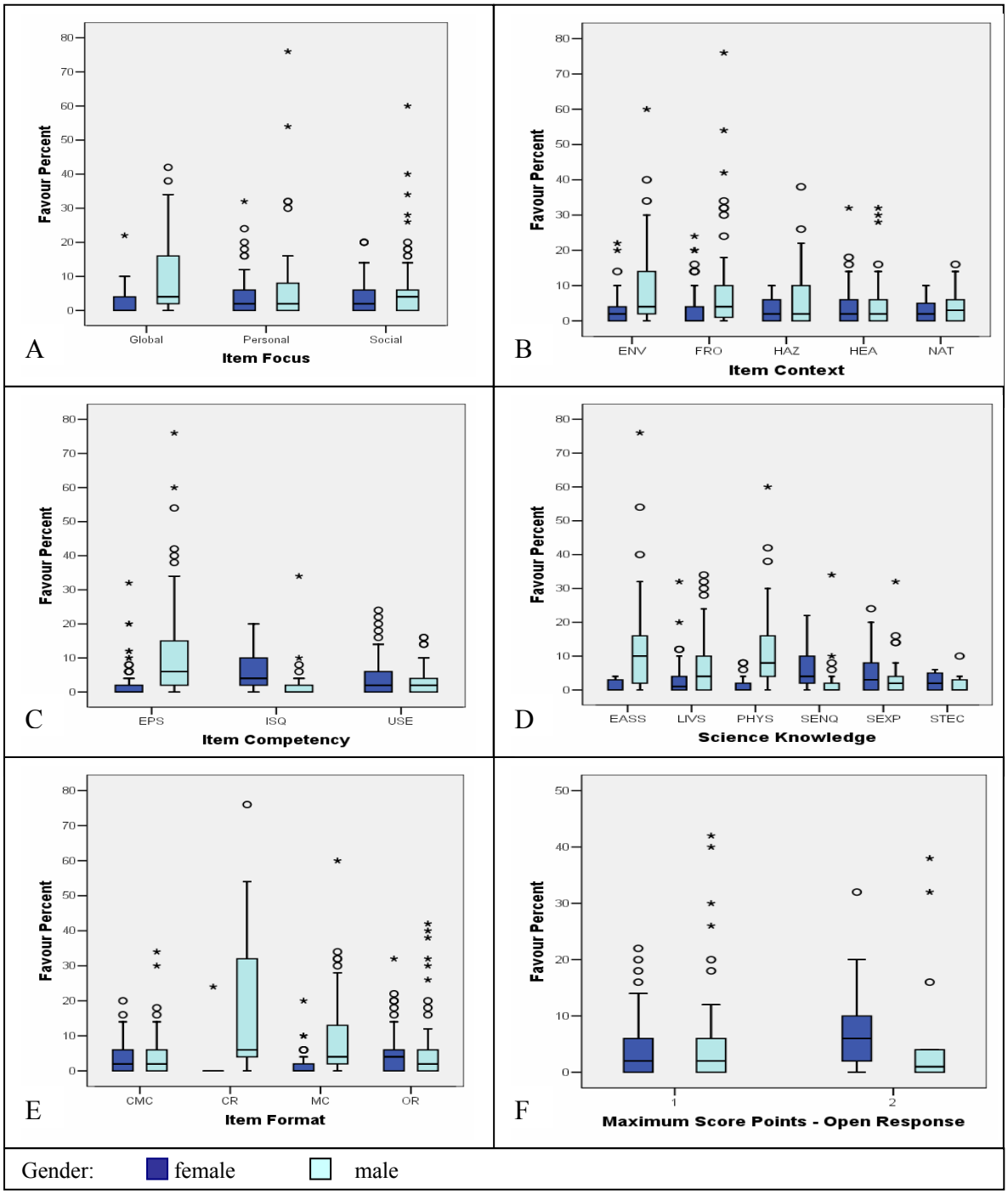
The mean of the gender DIF across countries and items is -0.05 with a SD of 0.25, and the mean of the unsigned DIF (absolute value of the DIF) is 0.33 with a SD of 0.12. On average, about 10% of the 210 items (or about 21 items) are flagged by each country (6.7% favour males, 3.3% favour females), in particular 12% flagged by OECD countries (7.9% favour males, 4.1% favour females). The correlations between item difficulty estimates by male and female groups vary from 0.87 to 0.96 with a mean of 0.92. This consistency is higher than those in the countries or languages analysed in the previous sections.

a) Gender DIF and item difficulty

The correlations between item difficulty estimates by a country group and its gender DIF are from -0.37 to 0.20 with a mean of -0.14. Among that, 54% of them are significantly negative correlations at the 0.05 level. Only four countries (8%) have a positive correlation, and among them one country, Japan, has a significantly negative correlation at the 0.05 level (r=0.20). This suggests a possible trend that with a similar ability level, males seem to perform relatively better than females on more difficult items.

b) Gender DIF and item characteristics

Figure 3 presents box plots of percentages of DIF flags that favour males or females separately across countries and with a reference by the item characteristics.



(Circles - outliers, stars –extreme cases and dash line – the median reference)

Figure 3. Gender DIF and Item Characteristics

- *Focus*: Males tend to perform better than females on "Global" items and slightly better than females on "Social" items, while in "Personal" there is not much difference between them (and with 9% DIF flags).

- *Context*: NAT, HEA and HAZ contain small percentages of gender DIF flags (8%, 10% and 11% respectively) and there is not a clear favour direction here. In contrast, ENV and FRO contain higher percentages of DIF flags (15% and 14%, respectively). Males tend to perform better on ENV and FRO items.

- *Competency*: ESP contains a higher percentage of DIF flags (17%) than do the other two categories (7% each). Males seem to have advantages in EPS items while females have advantages on ISQ items.

- *Scientific knowledge*: "knowledge of science" items tend to have more gender DIF than "knowledge about science" items (EASS: 15%, LIVS: 10% and PHYS: 12% compared with SENQ: 8%, SEXP: 9% and STEC: 3%). Males seem to perform relatively better on "Knowledge of science" items, while females perform relatively better on "Knowledge about science" items.

- *Item format*: CR items have the highest of gender DIF flags at 25%, compared with CMC: 8%, MC: 10% and OR: 10%. MC and CR items tend to favour males while females are slightly favoured on OR items. The picture is not clear on CMC items. Within OR items, females tend to perform relatively better than males on items with 2-score points.

c) Gender DIF and country feedback ratings

There is no significant correlation found at the 0.05 level (the correlations tend to be close to 0) between the number of gender DIF flags and the rating variables: Relevance, Interest, Authenticity, Culture, and Priority. But gender DIF only correlates significantly with Curriculum (r=0.21, p=0.007).

DISCUSION AND SUMMARY

The findings in this study are helpful for test development work as well as for interpreting test results of international comparable studies on science, or other cognitive domains. The design of this study can be applied for individual countries to find relative strengths and weakness of their student group or different subgroups.

DIF is not equivalent to bias. But DIF is unavoidable in international tests, especially when they include large, heterogeneous samples like in PISA. Items flagged with DIF may not be the problem but rather only a symptom of differences by factors such as culture, language, curriculum, and so on. Moreover, with cognitive tests, although one main dimension is defined, the variation of different specific ability or skills measured by different test items could be another factor causing the DIFs. This, however, does not support the notion that different constructs are measured.

In summary, the main findings are:

- DIF by gender across countries was found to be lower than DIF by countries or test languages. The latter DIFs may be caused more by factors such as heterogeneous sample, test translation, culture and curriculum.
- *Item focus*: Global items seem to be relatively easier for males than for females as well as for the five language test groups other than for English. The Arabic and Russian groups may be a bit disadvantaged on Personal items. The stability of item difficulty by countries looks equivalent among the three categories.
- *Context*: Males tend to perform relatively better on ENV and FRO items. FRO items seem to be relatively harder for the Arabic group (than English), but HAZ items seem to be relatively easier for them and the German and Spanish groups. NAT items seem have lower variation across the countries than other items.
- *Competency*: EPS Items appear to favour males, while ISQ items appear favour females. EPS items tend to favour the 5 test language groups other than, while ISQ items seem to be relatively harder for Arabic, German and Russian groups; and USE items tend to be relatively harder for the Arabic group. Additionally, the difficulties of USE items are more stable across the countries than that of items in other categories.
- *Scientific knowledge*: "knowledge of science" items (EASS, LIVS and PHYS) tend to favour males, while "knowledge about science" items (SENQ, SEXP and STEC) tend to favour females. PHYS items seem to be relatively easier for the five language groups other than English, while SENQ items seem relatively harder for the Arabic, Russian and German groups. Also the Spanish group found STEC relatively easier. Related to the country DIF, item difficulty seems to be most stable for SEXP items, and then for SENQ, while it is least stable for STEC items. The high percentages of flagged items in STEC by countries and language DIFs indicate that probably these items involve more curriculum or cultural problems than other items.
- *Item format*: MC and CR items seem to favour males, while OR items seem slightly to favour females. Moreover, MC items seem relatively easier for the English group, while the other five language groups seem to find OR items relatively easier. Compared with the English group, CR items tend to be relatively easier for French, Russian and Spanish, but they tend to be relatively harder for Arabic and German. The difficulties of OR items are less stable across countries than items of other formats. One possible explanation for this is that most of the OR items required markers which may not easily to be controlled equivalently in different countries.
- *OR – maximal score points*: Items with 2-score points seems to favour females. Moreover, it seems that the difficulty of 2-score point items is more stable by countries and test languages than those of 1-score point items. Additionally, the Arabic, Russian and German groups tend to relatively perform better on 1-score point items than 2-score point items.
- *DIF and country feedback ratings*: In general no relationship between these has been found yet, except that gender DIF correlates slightly with

the curriculum rating which indicates a possible clue that close curriculum items tend to favour females.

- *DIF and item difficulty*: The correlation between item difficulty estimates and gender DIF in this study suggests a trend that with a similar ability level, males seem to perform better than females on more difficult items. The correlations between country DIF or language DIF and item difficulty are still too small to give the relationship between them.

There are some possible limitations in this study which can indicate a direction for further research. This study is an exploratory investigation. The goal was to be inclusive or suggestive rather than confirmative. Additionally, there would be a problem of overlapping between the defined variables of item characteristics or, for example, item format with item focus, or small number of item in a category. More research is needed to confirm the findings and to disentangle any confounding resulting from such overlapping.

Moreover, this study only focuses mainly DIF and six dimensions of item characteristics defined in the PISA Framework. Other dimensions can be investigated in item levels such as spatial/visual content, verbal loading and so on. A detail DIF analysis with a reference to item curriculum levels would be very useful for explaining for the similarities and differences of the results found in PISA ("real life" approach) and in TIMSS ("curriculum" based approach), which is also a very large international survey on Mathematics and Science.

Reference

Adams, R. J. (2004). *PISA 2003 Technical Report.* OECD. Paris.

Becker, B. J. (1989). Gender and science achievement: A reanalysis of studies from two meta-analyses. *Journal of Research in Science Teaching*, 26, 141-169.

Birnbaum, A. (1968) Some Latent trait models and their use in inferring an examinee's ability. In F. M. Lord and Novick (Eds*.) Statistical Theories of Mental Scores* (pp.397-472). Reading, MA: Addition-Wesley.

Bock, R. D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. *Psychometrika,* 46, 443-459.

Bolger, N., & Kellaghan, T. (1990). Method of measurement and gender differences in scholastic achievement. *Journal of Educational Measurement*, 27, 165-174.

Burkam, D. T., Lee, V. E., & Smerdon, B. A. (1997). Gender and science learning early in high school: Subject matter and laboratory experiences. *American Educational Research Journal*, 34, 297-331.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Newbury Park, CA: Sage.

Cole, N. S. (1997, May). *The ETS gender study: How males and females perform in educational settings*. Princeton, NJ: Educational Testing Service.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland and H. Wainer (Eds.) *Differential Item Functioning* (pp35-66). Hillsdale, NJ: Lawrence Erlbaum.

Dorans, N. J., & Kulick, E. (1983). *Assessing unexpected differential item performance of female candidates on SAT and TSWE forms administered in December 1977: An application of the standardization approach* (RR-83-9). Princeton, NJ: Educational Testing Service.

Dorans, N. J., & Kulick, E. (1983). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance of female candidates on the Scholastic Aptitude Test. *Journal of Educational Measurement*, 23, 355-368.

Embretson, S.E, & Reise, S.P. (2000). *Item Response Theory for Psychologists*. Lawrence Erlbaum: Mahwah, NJ.

Ercikan, K. (1998). Translation effects in international assessments. *International Journal of Educational Research*, 29, 543–553.

Ercikan, K. (1999, April). *Translation DIF on TIMSS*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montréal, Quebec, Canada.

Ercikan, K. (2002). Disentangling sources of differential item functioning in multilanguage assessments. *International Journal of Testing*, 4, 199–215.

Ercikan, K. Gierl, M. J., McCreith, T., Puhan, G., & Koh, K. (2004). Comparability of Bilingual Versions of Assessments: Sources of Incomparability of English and French Versions of Canada's National Achievement Tests. *Applied Measurement in Education*, 17, 301-321.

Gierl, M. J., & Khaliq, S. N. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests*. Journal of Educational Measurement*, 38, 164-187.

Halpern, D. F. (1992). *Sex differences in cognitive abilities* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hambleton, R. K., & Swaminthan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer-Nijhoff.
Hambleton, R. K. (1993). Translating achievement tests for use in cross-cultural studies. *European Journal of Psychological Assessment*, 9, 57–68.
Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment*, 10, 229–224.
Hamilton, L. S. (1999). Detecting gender-based differential item functioning on a constructed-response science test. *Applied Measurement in Education*, 12, 211–235.
Hamilton, L. S., & Snow, R. E. (1998). *Exploring differential item functioning on science achievement tests (CSE Tech. Rep. No. 483)*. Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.
Ironson, G. H. (1977). *A comparative study of several methods of assessing item bias.* Unpublished doctoral dissertation, University of Wisconsin.
Ironson, G. H. & Subkoviak, M. (1979). A comparison of several methods of assessing item bias. *Journal of Educational Measurement,* 16, 209-225.
Jovanovic, J., Solano-Flores, G., & Shavelson, R. J. (1994). Performance-based assessments: Will gender differences in science achievement be eliminated? *Education and Urban Society*, 26, 352-366.
Klieme, E. and Baumert, J. (2001). Identifying national cultures of mathematics education: Analysis of cognitive demands and different item functioning in TIMSS, *European Journal of Psychology and Education*, 19, 3, 385-402.
Lapointe, A.E., Mead, N.A., & Askew, J.M. (1992). *Learning mathematics*. Princeton, NJ: Educational Testing Service.
Linn, M. C., & Hyde, J. S. (1989). Gender, mathematics, and science. Educational Researcher, 18(8), 17-19, 22-27.
Lord, F. M. (1980). *Applications of item response theory to practical problems.* Hillsdale, NJ: Erlbaum.
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.
Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149-174.
Mazzeo, J., Schmitt, A. P., & Bleistein, C. A. (1993). *Sex-related performance differences on constructed-response and multiple-choice sections of Advanced Placement Examinations (College Board Report No. 92-7)*. New York: College Entrance Examination Board.
Merz, W. R., & Grossen, N. F. (1979, April). *An empirical investigation of six methods for examine test item bias.* Paper presented at the annul meeting of the National Council on Measurement in Education, San Francisco. (ERIC Document Reproduction Service NO. ED 178566).
Mullis, I.V.S., Martin, M.O., Gonzalez, E.J., & Chrostowski, S.J. (2004), TIMSS 2003 international mathematics report. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.
Raju, N. S. (1988). The area between two item characteristics curves. *Psychometrika,* 54*,* 495-502.
Raju, N. S. (1990). Determining the significance of estimated signed and

unsigned areas between two item response functions. *Applied Psychological Measurement,* 14, 197-207.

Rasch, G. (1960/1980). *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Danish Institute for Education Research, 1960. (expended edition, Chicago: The University of Chicago Press, 1980.)

Ruder, L. M., Getson, P. R., & Knight, D. L. (1980). A Monte Carlo comparison of seven biased item detection techniques. *Journal of Educational Measurement,* 17, 1-10.

Sireci, S. G., & Berberoglu, G. (2000). Using bilingual respondents to evaluate translated-adapted items. *Applied Measurement in Education*, 35, 229–259.

Shepard, L. A., Camilli, G., Averill, M. (1981). Comparison of procedures for detecting test-item bias with both internal and external ability criteria. *Journal of Educational Statistics, 6*, 317-375.

Shepard, L. A., Camilli, G., & Williams, D. M. (1985). Validity of approximation techniques for detecting item bias. *Journal of Educational Measurement,* 22, 77-105.

Subkoviak, M. J., Mack, J. S., Ironson, G. H., & Craig, R. D. (1984). Empirical comparison of selected item bias detection procedures with bias manipulation. *Journal of Educational Measurement,* 21, 49-58.

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.

Swaminathan, H., & Roger, H. J. (1990). Detecting Item Differential Functioning using logistic regression procedures. *Journal of Educational Measurement, 27*, 361-370.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35-113). Hillsadle, NJ: Lawrence Erlbaum.

Van der Vijver, F.,&Tanzer, N. K. (1998). Bias and equivalence in cross-cultural assessment. *European Review of Applied Psychology*, 47, 263–279.

Wright, B. D. and Stone, M. H. (1979). *Best Test Design*. Chicago: MESA Press.

Wright, B. D. and Masters, G. N. (1982). *Rating Scale Analysis*. Chicago: MESA Press.

Wu, M.L., Adams, R.J., and Wilson, M.R. (1997). *ConQuest: Multi-Aspect Test Software.* Camberwell: Australian Council for Educational Research.

Young, D. J., & Fraser, B. J. (1994). Gender differences in science achievement: Do school effects make a difference? *Journal of Research in Science Teaching*, 31, 857-871.

Zenisky, A.L., Hambleton, R.K., & Robin, F. (2004). DIF Detection and Interpretation in Large-Scale Science Assessments: Informing Item Writing Practices. *Educational Assessment*, 9 (1-2), 61-78.

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

Appendix 1 – Additional Tables

Table A1. Country – gender frequency

Country name	Short name	Females	Per cent	Males	Per cent	Missing	Per cent	Total
Australia	AUS	1014	50.3%	1000	49.7%	0		2014
Austria	AUT	976	49.9%	980	50.1%	0		1956
Azerbaijan	AZE	624	47.8%	680	52.1%	1	0.1%	1305
Belgium	BEL	1117	52.5%	1009	47.5%	0		2126
Canada	CAN	751	51.1%	707	48.1%	11	0.7%	1469
Chinese Taipei	TWN	972	45.5%	1164	54.5%	0		2136
Colombia	COL	943	54.8%	778	45.2%	0		1721
Croatia	HRV	796	52.8%	712	47.2%	0		1508
Czech	CZE	695	50.8%	672	49.2%	0		1367
Denmark	DNK	939	48.8%	987	51.2%	0		1926
Estonia	EST	1013	47.0%	1144	53.0%	0		2157
Finland	FIN	599	49.8%	604	50.2%	0		1203
France	FRA	749	53.0%	665	47.0%	0		1414
Germany	DEU	2952	50.5%	2896	49.5%	0		5848
Greece	GRC	643	51.9%	595	48.1%	0		1238
Hongkong	HKG	582	48.2%	625	51.8%	0		1207
Hungary	HUN	665	51.8%	620	48.2%	0		1285
Iceland	ISL	595	51.4%	563	48.6%	0		1158
Indonesia	IDN	666	49.2%	688	50.8%	0		1354
Ireland	IRL	785	53.7%	676	46.3%	0		1461
Israel	ISR	1327	57.1%	999	42.9%	0		2326
Italy	ITA	854	45.3%	1032	54.7%	0		1886
Japan	JPN	642	48.7%	676	51.3%	0		1318
Jordan	JOR	785	53.4%	684	46.6%	0		1469
Kazakhstan	KGZ	1745	54.4%	1463	45.6%	0		3208
Korea	KOR	899	40.5%	1323	59.5%	0		2222
Latvia	LVA	717	50.4%	705	49.6%	0		1422
Lithuania	LTU	653	47.0%	735	53.0%	0		1388
Luxembourg	LUX	597	54.0%	509	46.0%	0		1106
Mexico	MEX	656	52.1%	604	47.9%	0		1260
Montenegro	QMN	651	48.0%	704	52.0%	0		1355
Netherlands	NLD	593	46.8%	674	53.2%	0		1267
New Zealand	NZL	717	61.2%	455	38.8%	0		1172
Norway	NOR	594	49.1%	617	50.9%	0		1211
Poland	POL	686	50.9%	662	49.1%	0		1348
Portugal	PRT	1022	56.1%	801	43.9%	0		1823
Qatar	QAT	722	48.5%	722	48.5%	45	3.0%	1489
Romania	ROU	820	50.5%	805	49.5%	0		1625
Russia	RUS	949	56.0%	747	44.0%	0		1696
Scotland	QSC	238	47.8%	256	51.4%	4	0.8%	498
Serbia	QSB	693	49.0%	721	51.0%	0		1414
Slovak Republic	SVK	1014	53.8%	871	46.2%	0		1885
Slovenia	SVN	752	47.4%	836	52.6%	0		1588
Spain	ESP	1990	50.8%	1924	49.2%	0		3914
Sweden	SWE	602	49.1%	623	50.9%	0		1225

Table A1. Country – Gender Frequency (*continue*)

Country name	Short name	Females	Per cent	Males	Per cent	Missing	Per cent	Total
Switzerland	CHE	1144	50.5%	1120	49.5%	0		2264
Tunisia	TUN	636	52.0%	588	48.0%	0		1224
Turkey	TUR	648	43.0%	860	57.0%	0		1508
United Kingdom	QUK	540	51.8%	503	48.2%	0		1043
United States	USA	653	53.3%	573	46.7%	0		1226
Total		42615	50.6%	41557	49.3%	61	0.1%	**84233**

Table A2. Test Language Frequency

Language Name	Short name	Frequency	Percent
Arabic	ARA	4814	13.4
English	ENG	8686	24.2
French	FRE	3139	8.7
German	GER	10877	30.3
Russian	RUS	3574	10.0
Spanish	SPA	4809	13.4
Total		35899	100.0