European Educational Research Journal
Volume 8 Number 1 2009
www.wwwords.eu/EERJ

Gender Gap in Comparative Studies of Reading Comprehension: to what extent do the test characteristics make a difference?

DOMINIQUE LAFONTAINE & CHRISTIAN MONSEUR
University of Liège, Belgium

ABSTRACT In this article we discuss how apparently indicators that may appear straightforward, such as gender differences, need to be interpreted with extreme care. In particular, we consider how the assessment framework, and the methodology of international surveys, may have a potential impact on the results and on the indicators. Through analysis of Programme for International Student Assessment (PISA) data we show how increases or decreases in the achievement of some groups of students (either of whole countries or population subgroups like males and females) can, at least partially, result from variations in the framework or the methodology of the respective assessments. The analyses provide evidence that the gender gap is larger for open-ended questions, for continuous texts and for more cognitively demanding reading tasks.

Introduction

It is current practice to build efficiency and equity indicators of education systems based on national or international surveys. International agencies like the Organisation for Economic Cooperation and Development (OECD), the European Union (EU) or the International Association for the Evaluation of Student Achievement (IEA) regularly release updated sets of indicators, for instance, *Education at a Glance* (OECD, 2007) and *Key Data on Education in the European Union* (European Commission et al, 2005). Among equity indicators, differences of achievement in various domains (but mainly reading, mathematics and science) between males and females are displayed in each set of indicators. Computing those indicators does not raise technical problems and there is relative consensus among modern democratic societies that the gender achievement gap should be reduced. Until recently, the major concern was to improve females' achievement in scientific domains; currently, there is a growing concern about males' underachievement in reading literacy.

The release of such indicators largely informs discussion in the public audience and is critical for the monitoring of educational systems. Those sets of indicators undoubtedly provide useful information for decision makers. However, the information they provide is condensed, often weakly contextualised and the product (the indicator) could unduly be attributed an absolute value, though its scope is in fact limited: on the one hand, to the study it is based upon; on the other hand, to the technical processes that led to its construction. Even among experts and *a fortiori* among policy makers and public audience, there is a serious risk of over-interpreting indicators.

There is a growing interest in trends indicators, and international surveys in education are thus conducted on a regular basis. Educators or policy makers might be eager to compare the results between successive assessments and to interpret the differences in terms of evolution – increase or decrease. For instance, the Programme for International Student Assessment (PISA) 2003 international report (OECD, 2004) presents in a figure the gender difference on the combined

 http://dx.doi.org/10.2304/eerj.2009.8.1.69

reading literacy scale for the 2000 and 2003 assessments. One might also compare the gender differences between the IEA Reading Literacy Study (IEARLS) and the PISA 2000 study.

Objectives of the Study

The aim of the present study is to explore the impact of some of the test characteristics, especially the question format, the reading process and the type of text, on gender equity indicators in reading literacy comparative assessments. The starting point for this research is the inconsistencies in the gender equity indicators between the 1991 IEARLS and the PISA 2000 study. In 1991, the gender gap in IEA reading comprehension among 14 year-olds was rather limited (7 score points on average on a scale with an international standard deviation of 100) and statistically non-significant in many countries (Elley, 1994). About 10 years later (PISA 2000), the gender gap is very much larger (32 score points on a comparable scale) (OECD, 2001). To what extent is this apparent increase in the achievement gap between males and females 'true' and to what extent does it reflect the technical parameters of both studies?

While assessing the same domain or the same latent variables in psychometric terms (reading literacy), the two studies certainly differ on several crucial aspects, among others definition of the population, types of stimulus, balance of different question formats, reading processes assessed and so on. Among those numerous potentially influential factors that might affect the gender equity indicator, this article focuses on some crucial parameters of the assessment framework, in particular on the question format, the reading process and the type of text.

Data from the Reading Literacy Study led by the IEA in 1991 (Elley, 1994) and from PISA, led by the OECD in 2000 (OECD, 1999, 2001; Adams & Wu, 2002) will be used. Most of the statistical analyses will be based on the PISA data; however, the descriptive elements needed for comparisons between the two studies are provided below.

Comparison of the IEA Reading Literacy Study and the PISA 2000 Study

The IEA Reading Literacy Study was conducted in 1991 in 31 educational systems and aimed at assessing reading comprehension in the two grades attended by most 9 and 14 year-olds. For the present study, only data from the population of 14 year-olds will be used. The 14-year-olds' test comprised 89 items all included in a single booklet administered to all sampled students.

The PISA 2000 study was implemented in 32 countries and assesses reading literacy amongst 15 year-olds. The reading assessment has 129 different items, rotated in 9 different booklets (for details about the test design, see Adams & Wu, 2002). In this study, the results of Liechtenstein will not be included, as the sample size was only about 350 students.

The main aspects of the two studies are presented side by side and summarised in Table I.

Population Definition

The IEARLS and the OECD PISA 2000 study differ in their target population. Students in PISA 2000 are somewhat older, but the grade population versus the age population constitutes the main difference. This has no or little consequences for education systems with no or low grade repetition rates, but it makes a difference for education systems with high rates of grade repetition. In PISA, grade repeaters will attend a lower grade and the grade attended has a major impact on achievement (Kirsch et al, 2003).

Furthermore, Monseur & Lafontaine (2006) have shown that the choice of the population definition has a limited but significant impact on the width of the gender gap, as boys more often repeat a grade than girls. For instance, in the French-speaking community of Belgium, more than 60% of the females are in the expected grade, i.e. 10th Grade in this example, but less than 50% of the males are attending the expected grade. Due to the dependency between gender and grade retention, the gender gap is increased with an age population. While in Belgium the difference between boys and girls is about 32 point in the PISA 2000 reading scale, within a particular grade, it is equal to 19 points on average.

Stimulus for the Reading Tasks

Both studies used continuous and non-continuous texts as reading stimulus, and a variety of types of texts. Differences between the two studies are slight at this level. There are no argumentative texts in IEARLS, but the effect of this on the relative performances of males and females is not known. Items based on narrative stimulus are less frequent in PISA 2000, and, everything being equal, it could somewhat reduce the gap between males and females as typically females read narrative texts more often than males, so they could be more familiar with the narrative texts. For instance, in PISA 2000, 15 year-olds were asked to report on the types of text they usually read:

> Males report more frequently than females that they mainly read newspapers, magazines and comics rather than books (especially fiction). ... Conversely, across all countries, females ... identify themselves as reading newspapers, magazines, books (especially fiction) but not comics. (Kirsch et al, 2003, pp. 116-117)

But many other aspects can counteract this influence – for example, content of the text, reading processes assessed, question format – consequently, it is difficult to assess exactly what impact this slight difference might have on the gender achievement gap.

Reading Aspects Assessed

In both studies, IEARLS and PISA, several reading aspects or processes are assessed. The items from IEARLS have been reclassified according to the categories used in the PISA 2000 framework. The categories used in the IEARLS were somewhat different (Elley, 1994). For retrieving and interpreting, the proportions of items are more or less equivalent. No item is aimed at assessing the aspect 'reflect upon the text' in IEARLS.

Question Format

One of the most striking differences between the studies is the relative proportion of multiple-choice and open-ended questions. All IEARLS items have an 'objective' answer (multiple-choice or short answers which could be scored 'correct' or 'incorrect' without any interpretation from the markers). In PISA 2000, almost half of the questions are constructed as open-ended and the scoring relies on detailed correction procedures.

Target population	Population definition	IEA Reading Literacy (1991) Grade attended by the majority of 14 year-olds %	PISA 2000 15 year-olds, regardless of the grade attended %
Stimulus for the reading task	Continuous texts:		
	narrative	33	13
	expository/descriptive	29	35
	argumentative/injunctive	0	18
	Non-continuous texts	38	34
Reading aspects assessed	Retrieving or locating information (literal or paraphrase)	42	30
	Interpreting (inferring, finding the main idea)	58	50
	Reflecting and evaluation	0	20
Question format	Multiple-choice	75	45
	Open-ended short answer	22	11
	Constructed open-ended	3	45

Table I. Differences between the IEA Reading Literacy Study and PISA 2000.

Review of the Literature

Extensive research has been dedicated to the effect of item format on achievement of males and females, including numerous studies attempting to address the question of test bias. According to a synthesis carried out by Bennett (1993), 'several studies have found that relative to males, females perform better on constructed-response than on multiple-choice items' (p. 20) and 'studies reviewed by Traub & MacRury, 1990, also support this finding' (Bennett, 1993, p. 20).

While much research has supported this general finding (Mazzeo et al, 1991; DeMars, 2000), other studies have shown that this pattern does not hold true in all subject areas. Mullis et al (2000) have analysed gender differences by item format in IEA's TIMSS (Third International Mathematics and Science Study). There were three different types of item format: multiple-choice, short answer and extended response. There were few significant differences: almost no significant difference for science and maths at Grade 4, and few differences at Grade 8; most of the differences were observed in the final year of secondary school. According to Mullis et al, 'results were not consistent across grades or subject areas, although there was a slight tendency at the upper grades for males to have outperformed females in more countries on free-response mathematics items and on multiple-choice science items' (Mullis et al, 2000, p. 98).

Similarly, Routitsky & Turner, analysing PISA 2003 field trial data – mathematics items for a population of 15 year-olds in 42 countries – found mixed and nuanced results regarding the interaction between item format and gender:

> Preliminary indications are that extended open-constructed response may favour girls and short answer questions may favour boys. However, as the item difficulty increases, the likelihood to favour boys for both open constructed response and short answer items increases. (Routitsky & Turner, 2003, p. 25)

In addition, analyses conducted across all countries show that 'students of lower ability across all countries are on average doing better on the multiple-choice items than on both extended open constructed response and short answer items' (Routitsky & Turner, 2003, p. 26). This last finding could explain why the interaction between gender and item format is nearly always observed in subject areas in which girls traditionally outperform boys (notably, reading and writing) and that results are less conclusive for subject areas in which boys traditionally outperform girls (mathematics and science).

This finding draws attention to the 'considerable potential for interaction effects' (Bennett, 1993, p. 23) as far as item format is concerned. The general pattern of interaction with gender could be modified depending on the subject matter, the item difficulty or the student's ability and even other aspects like content or cognitive process assessed.

Hypotheses

Firstly, according to the literature, we hypothesise that, as far as reading ability is concerned, the gender gap will be larger for constructed open-ended questions than for multiple-choice questions, in favour of females (hypothesis 1).

Secondly, as stated by Bennett (1993), there is a 'considerable potential for interaction effects' (p. 23); it is therefore hypothesised that additional interaction effects may be observed. The gender gap for multiple-choice item and the gender gap for open-ended item could vary from a reading aspect (retrieve, interpret, reflect upon the text) to another one. As questions assessing reflection are cognitively more demanding than locate or interpret questions, the increase in the gender gap will be larger for the former aspect than for the latter (hypothesis 2).

Thirdly, the impact of question format may also differ according to the type of texts (continuous/non-continuous). One could assume that the gender gap for the different question formats will be more important for aspects in which males generally perform at a lower level then females. Consequently, we hypothesise that the gender gap according to question format will be larger for continuous than for non-continuous texts (hypothesis 3).

Analyses

The PISA 2000 released international database contains five subscales for reading literacy: (i) reading retrieve information, (ii) reading interpret, (iii) reading reflect, (iv) reading continuous texts and (v) reading non-continuous texts. The mixed coefficients parameters multinomial logit model as described by Adams et al (1997) was used to scale the PISA data and implemented by Conquest software (Wu et al, 1997). This model is a generalised form of the Rasch model. For more details see Adams & Wu (2002).

For the purpose of this article, 10 new reading subscales were generated according to the same model with, however, a few differences that have no impact on the results presented in the article:

1. The PISA 2000 initial subscales were generated according to a multidimensional model, i.e. the three reading aspect subscales were generated simultaneously in conjunction with the mathematics and science scales. The same model was used for the generation of the two text type subscales. This multidimensional model is required for obtaining unbiased estimates of correlations between scales. As such relationships are not part of this study, the 10 new subscales were generated according to a one-dimensional model.
2. To provide unbiased estimates of the relationships between contextual data and performance, the PISA 2000 initial subscales were conditioned on all student-level background variables. In this study, the contextual information used for conditioning was limited to gender. However, the conditioning also included the booklet identification for counterbalancing the booklet effect observed on the PISA 2000 data. For more information, see Adams & Wu (2002).

As for the PISA 2000 initial reading subscales, the combined reading literacy equation was used to transform the logit score on the PISA reading scale (with a mean of 500 and a standard deviation of 100).

The generation of Plausible Values was implemented at the country level.

The five subscales presented above were rebuilt but per item type, i.e. multiple-choice item and open-ended items. This ends up with 10 subscales. In other words, the 10 new subscales simply split the five initial subscales by item format. Tables II and III present, by subscale, the number of items.

	Multiple-choice items	Open-ended items	Total
Reading/retrieve information	12	24	36
	(16.7)	(19.3)	
Reading/interpret	43	21	64
	(29.8)	(34.2)	
Reading/reflect	5	24	29
	(13.5)	(15.5)	
Total	60	69	129

Table II. Distribution of the PISA 2000 reading items by process and by item format and expected frequencies (in brackets).

	Multiple-choice items	Open-ended items	Total
Reading/Continuous	45	42	87
	(40.5)	(46.5)	
Reading/Non-continuous	15	27	42
	(19.5)	(22.5)	
Reading/Total	60	69	129

Table III. Distribution of the PISA 2000 reading items by text type and by item format.

A chi-square test was performed on the two distributions of items, i.e. item format per reading process and item format per text type. The independence test between item format and reading

process is rejected ($p < .001$) but the independence test between item format and text type is not rejected ($p = .09$).

The distribution of the PISA 2000 items between reading aspects and question format is not balanced, as shown by Table III and its associated chi-square. This unbalanced design therefore confounds the effect of item format and of reading aspect on the gender difference. In other words, if a higher gender gap in the reflecting subscale were observed, then it could not be directly interpreted as an effect of the assessed reading aspect, because there are more open-ended items for that aspect.

The numbers presented in brackets represent the expected number of items that would allow the interpretation of the differences of the gender gap between subscales as an effect of the reading aspect. The comparison between observed and expected numbers of items allows location of the imbalances. For the aspect 'interpret', multiple-choice items are proportionally more numerous (43 versus 29.8) than open-ended questions (21 versus 34.2). For the aspect 'retrieve information', the reverse is true: open-ended questions (24 versus 19.3) are proportionally more numerous than multiple-choice items (12 versus 16.7). As far as the aspect 'reflect' is concerned, open-ended questions are far more frequent (24 versus 15.5) than multiple-choice items (5 items versus 13.5). The imbalance, in this last case, is due to the extreme difficulty of writing closed questions that assess these specific skills.

It is therefore crucial to estimate the relative impact of question format and reading aspect on the gender gap achievement, which results in testing the hypothesis of an interaction between question format and reading aspect.

The comparison between the observed and the expected distribution of items per item format and per type of texts also shows some imbalances but quite a lot smaller than for the reading aspects.

Results

Before presenting the results of the analyses, the difference in reading proficiency between males and females in IEARLS 1991 and PISA 2000 will be briefly presented.

In IEARLS, on average, girls outperform boys by 7 points ($p < .05$) on a scale means 500, standard deviation 100. In 13 countries out of the 31, the gender difference is not significant ($p > .05$). There is no difference for non-continuous texts. The difference between boys and girls is equal to 3 and 18 respectively for informative texts and for narrative texts. The difference is therefore larger for narrative texts. No subscales are available for reading aspects/processes.

Table IV presents the standardised gender differences on the PISA 2000 subscales (SD being equal to 1). As the international standard deviation was different for each subscale, the standardised difference has been preferred.

Reading subscale	Standardised gender difference
Combined scale	0.32 (0.01)
Retrieve information	0.23 (0.01)
Interpret	0.28 (0.01)
Reflect	0.41 (0.01)
Continuous texts	0.39 (0.01)
Non-continuous texts	0.17 (0.01)

Table IV. Standardised gender difference in PISA 2000 (standard error in brackets) (OECD, 2001).

In PISA 2000, the average difference between males and females is 32 score points; the difference is statistically significant in every participating country and ranges from 14 score points (in Korea) to 53 score points (in Latvia). The gap between males and females is larger (in favour of females) for continuous texts, as in IEARLS; and larger for the aspect 'reflect upon the text' than for 'interpret the text' and 'retrieve information'.

The gender gap is obviously larger in PISA 2000 than in IEARLS (1991). To what extent can this 'growing' gap in reading proficiency between males and females be explained by influential parameters of the framework, namely the respective proportions of multiple-choice and open-ended questions?

Let us turn now to the results of those analyses. In all countries, the standardised gender difference is higher for the open-ended items than for the multiple-choice items. The median of these standardised gender differences is respectively 0.20 and 0.28 for the multiple-choice item scale and for the open-ended scale. The first hypothesis is thus confirmed, showing that the gender gap in favour of females is larger for open-ended items than for multiple-choice items.

However, there are large differences in the country profiles. The relative increase of the standardised difference (i.e. the standardised difference for multiple-choice items divided by the standardised difference for open-ended items) between males and females when shifting from all multiple-choice to all open-ended questions varies between 14% (in Portugal) and 114% (in Korea). On average among participating countries, it reaches 52%. To put it another way, moving from an assessment of reading comprehension with 100% multiple-choice items to an assessment balancing multiple-choice and open-ended items obviously will have an impact on the width of the gap between males and females.

However, due to the lack of independency between item formats and reading aspects in PISA 2000, further investigations are needed. We have also seen in Table IV that the PISA gender gap is larger for the aspect 'reflect' than it is for retrieving information or for interpreting. Unfortunately, no question in IEARLS specifically addressed this aspect, another difference in the framework which could account for a larger gender gap in PISA.

	Retrieve		Interpret		Reflect	
	MCQ	Open-ended	MCQ	Open-ended	MCQ	Open-ended
Mean	19	29	29	35	32	47
	(0.95)	(0.99)	(0.86)	(0.92)	(1.12)	(0.93)
Mean (standardised)	0.19	0.26	0.30	0.35	0.28	0.47
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)

Table V. Mean of gender differences per reading aspect and per item format. Standard errors are provided in brackets.

As shown by Table V, the gender differences vary according to the item format and according to the reading aspect. As expected, the smaller difference is associated with the retrieve aspect assessed only by multiple-choice items and the highest is associated with the reflect aspect assessed only by open-ended items. The influence of reading aspect and item format on gender difference is quite substantial as it can range from a standard deviation of 0.19 to 0.47. These results confirm the second hypothesis, i.e. the impact of question format on the gender gap is higher for the aspect 'reflect' than for 'interpreting the text' and for 'retrieving information'. Although at first glance Table V results would suggest that item type has a stronger influence on the 'reflecting' aspect than on the two other aspects, such an interpretation would be tenuous as the 'reflect / multiple-choice' scale only consists of five items.

	Continuous texts		Non-continuous texts	
	MCQ	Open-ended	MCQ	Open-ended
Mean	33	48	12	21
	(0.89)	(0.94)	(1.05)	(0.99)
Mean (standardised)	0.34	0.45	0.10	0.19
	(0.01)	(0.01)	(0.01)	(0.01)

Table VI. Mean of gender differences by text type and by item format. Standard errors are provided in brackets.

Table VI presents the mean of the gender difference and its standardised equivalent by text type and by item format. Table VI results confirm that the gender gap according to question format is

higher for continuous than for non-continuous texts. However, in that case, there is no interaction as the impact of question format on gender gap is more or less equivalent for continuous and for non-continuous texts.

The previous analyses have demonstrated the impact of text types, reading aspects and question format on the gender gap. Further, the IEARLS and the PISA 2000 international reports have shown a large variability of the gender gap between countries. To estimate the relative impact of the different factors at work, a variance decomposition of the gender differences has been performed on the six reading aspect subscales and on the four text type subscales. The analyses had to be performed independently because there would have been too few items per cell if the aspect, the text type and the item format had been crossed. In both analyses, the dependent variable was the non-standardised gender difference. The first analysis includes as independent variables reading aspect and item format. The second analysis includes the text type and the item format. The country was also added to both analyses for controlling that potentially influential source of variation.

Effect	% of variance	Effect	% of variance
Country (cnt)	40	Country (cnt)	21
Process (pro)	24	Text (tex)	52
Question format (ques)	17	Question format (ques)	12
Cnt * pro	0	Cnt * tex	4
Cnt * ques	5	Cnt * ques	3
Pro * ques	5	Tex * ques	2
Cnt * pro * ques	9	Cnt * tex * ques	3

Table VII. Variance decomposition of the gender differences.

First of all, even if the country effect substantially varies between the two analyses (respectively 40% and 21% of the total variance), it remains in both cases below 50%. In other words, the gender difference is not only a country characteristic; it also depends to a large extent on test characteristics and their potential interactions with the country variable. It does prove that this equity indicator should not be interpreted in absolute terms but it is related to the assessment framework. Concretely, one can manipulate the size of the gender gap, depending on what is included in the test.

The text type appears to have the larger effect on the gender differences. More than 50% of the variance is attributable to the text type. In other words, the best way to maximise the gender gap is to only include continuous texts, while including only non-continuous text would minimise it. The reading aspects also substantially influence the size of the gender gap. The item format and its associated interactions explain about 20-30% of the gender differences observed in the countries.

Finally, the different interactions that involve the countries count in both variance decompositions for less than 15%. It means, on average, that if a country presents a substantial gender gap for one subscale, the gender gap will also be substantial for the other scales and vice versa.

Conclusions

The aim of the present study was to explore the impact of some of the test characteristics, especially the question format, the reading process and the type of text, on gender gap in reading literacy comparative assessments. The hypothesis of an interaction between item format and gender is supported by the data in each of the 31 participating countries: the gap in reading proficiency between males and females is larger for open-ended than for multiple-choice items. This finding is congruent with the literature review.

The hypothesis of a modulation of the pattern of interaction between gender and item format by the reading aspect assessed is also partially supported by the data. On average among countries, the impact of question format will be larger for the aspect 'reflect upon the text' than for 'interpret the text' and 'retrieve information'. This finding nevertheless has serious limitations and should be

regarded with caution, due to the small number of items in some of the cells (five multiple-choice items only for the aspect 'reflect').

The variance analysis carried out in order to disentangle the effect of item format from other confounded variables (mainly the reading aspect) clearly shows that the reading aspect has a larger impact (24% of variance explained) than item format on the difference in reading achievement between males and females. But item format also makes a striking difference (16% of variance explained).

Also, the type of text appears to be one of the major factors contributing to gender differences (53% of variance explained). This result is not surprising and can be related to the differences in written material regularly read by males and females respectively.

Coming back to the initial question behind this study – has the gender gap grown between 1991 and 2000 or is it an artefact? – one can argue on the basis of the findings that the decrease of the proportion of multiple-choice items between IEARLS and PISA has potentially influenced the growth of the gender gap. On average, a test composed of 100% of open-ended items will lead to a gender gap 53.6% larger than a test including only multiple-choice items. The impact of item format on gender differences is worth considering.

Moreover, another parameter of the framework – the reading aspects assessed – also accounts for about a quarter of the variance of the gender gap achievement. But the type of texts especially affects the gender gap as it accounts for more than half of the variance of the gender gap. Additional research is now needed to explore in more depth the reasons for the apparent growth of the gender gap in reading proficiency between the early 1990s and 2000.

Furthermore, Monseur & Lafontaine (2006) have shown that the choice of the population definition in terms of grade versus age also has a limited but significant impact on the width of the gender gap in educational systems with high retention rates. Taken together, those various methodological choices constituting the framework for the assessment influence to quite a large extent the width of the achievement gap in reading comprehension between males and females. Indicators of gender equity based upon assessments which have made different methodological choices are not obviously comparable. Before considering that the gender gap noticed in PISA is of serious concern, one has to consider carefully the nature of the reading tasks administered to the students. Another reading assessment, assessing different tasks, with different stimulus and/or different item format could have led to quite divergent conclusions on the respective reading proficiencies of males and females.

Recently, the organisations (IEA, OECD) in charge of international comparative assessments have come to a turning point, moving from an agenda based on isolated surveys to an agenda aimed at measuring trends through repeated cycles (PISA, TIMSS-R, PIRLS). This new perspective opens the way for truly comparable assessments and no doubt constitutes substantial progress in monitoring education systems on reliable grounds. However, the reading experts and test developers in charge of those assessments should be careful to guarantee a similar balance of the various components of the reading framework in successive assessments; otherwise the validity of the trends indicator is likely to be jeopardised. If, for instance, the PISA 2009 reading were to include more 'reflect' items or, especially, more continuous texts or more open-ended items, it would clearly result in an increased gender gap compared to the PISA 2000 assessment. Of course, some powerful considerations or arguments could or should lead to revision of the framework. In conclusion, we want to stress the importance of carefully arbitrating the advantages and disadvantages of changes and adaptations between successive assessments. Weakly founded modifications should definitely be avoided.

References

Adams, R.J., Wilson, M.R. & Wang, W. (1997) The Multidimensional Random Coefficients Multinomial Logit Model, *Applied Psychological Measurement*, 21, 1-24. http://dx.doi.org/10.1177/0146621697211001

Adams, R.J. & Wu, M. (Eds) (2002) *PISA 2000 Technical Report*. Paris: Organisation for Economic Cooperation and Development.

Bennett, R.E. (1993) On the Meaning of Constructed Response, in R.E. Bennett & W.C. Ward (Eds) *Construction versus Choice in Cognitive Measurement. Issues in Constructed Response, Performance Testing, and Portfolio Assessment*, 1-27. Hillsdale: Lawrence Erlbaum Associates.

DeMars, C.E. (2000) Test Stakes and Item Format Interaction, *Applied Measurement in Education*, 13(1), 55-77. http://dx.doi.org/10.1207/s15324818ame1301_3

Elley, W.B. (1994) *The IEA Study of Reading Literacy: achievement and instruction in thirty-two school systems*. London: Pergamon.

European Commission, Eurydice & Eurostat (2005) *Key Data on Education in the European Union*. Brussels: European Commission, Eurydice & Eurostat.

Kirsch, I., de Jong, J., Lafontaine, D., et al (2003) *Reading for a Change. Performances and Engagement. Results from PISA 2000*. Paris: Organisation for Economic Cooperation and Development.

Mazzeo, J., Schmitt, A.P. & Bleistein, C.A. (1991) Do Women Perform Better, Relative to Men, on Constructed-Response Tests or Multiple-Choice Tests? Evidence from the Advanced Placement Examinations. Paper presented at the annual meeting of the American Educational Research Association, Chicago, April.

Monseur, C. & Lafontaine, D. (2006) Le caractère relatif des indicateurs de tendance [Relative characteristics of trends indicators], *Revue suisse des Sciences de l'Education*, 3-28, 353-371. (In French)

Mullis, I.V.S., Martin, M.O., Fierros, E.G., Goldberg, A.L. & Stemler, S.E. (2000) *Gender Differences in Achievement. IEA's Third International Mathematics and Science Study.* Chestnut Hill, MA: TIMSS International Study Centre, Boston College.

Organisation for Economic Cooperation and Development (OECD) (1999) *Measuring Students' Knowledge and Skills. A New Assessment Framework. PISA 2000*. Paris: OECD.

Organisation for Economic Cooperation and Development (OECD) (2001) *Knowledge and Skills for Life: first Results from PISA 2000*. Paris: OECD.

Organisation for Economic Cooperation and Development (OECD) (2004) *Learning for Tomorrow's World: first results from PISA 2003*. Paris: OECD.

Organisation for Economic Cooperation and Development (OECD) (2005) *PISA 2003 Technical Report*. Paris: OECD.

Organisation for Economic Cooperation and Development (OECD) (2007) *Education at a Glance 2007. OECD Indicators*. Paris: OECD.

Routitsky, A. & Turner, R. (2003) Item Format Types and their Influences on Cross-National Comparisons of Student Performance. Paper presented at the annual meeting of the American Educational Research Association, Chicago, April.

Traub, R.E. & MacRury, K. (1990) Antwort-auswahl – vs freie-antwort-aufgaben bei lernerfolgs-test [Multiple choice vs. open ended for the learner test], in K. Ingenkamp & R.S. Jäger (Eds) *Tests und trends 8: Jahrbuch der pädagogishen diagnostik*, 128-159. Weinheim: Beltz Verlag. (In German)

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

DOMINIQUE LAFONTAINE is professor and head of the Department of Analysis of Systems and Practices in Education (aSPe) of the University of Liège, Belgium. She is in charge of several courses (analysis and monitoring of education systems, teaching and learning processes, experimental pedagogy, educational psychology). She has conducted numerous national large-scale assessments, especially in the field of reading. She has been national research coordinator for several comparative studies (IEA Reading Literacy and PISA). She is an expert consultant for national and international large-scale assessments (Member of the Reading Expert Group for PISA and for the IEA-PIRLS 2006) and member of the Board for monitoring compulsory education in the French Community of Belgium. *Correspondence*: Dominique Lafontaine, University of Liège, Boulevard du Rectorat 5, B32, B-4000 Liege, Belgium (dlafontaine@ulg.ac.be).

CHRISTIAN MONSEUR worked for 10 years as a research assistant at the University of Liège, Belgium. He participated in the IEA Computers in Education study and acted as Third International Mathematics and Science Study (TIMSS) National Research Coordinator for the French Community of Belgium. He was data manager for the OECD/PISA study in 2000 and project director for the PISA Plus study. As data manager for the PISA project, he performed all statistical analyses for the OECD initial report *Knowledge and Skills for Life* (OECD, 2001) and for the thematic report *Reading for a Change: performance and engagement across countries* (OECD, 2002). He

wrote the PISA 2003 database manual in collaboration with Keith Rust and Sheila Krawchuk. In 2005, he wrote his PhD thesis and became professor at the University of Liège in 2006. Dr Monseur has been a consultant to WESTAT and is a member of technical expert groups for the IEA ICCS, the OECD PISA, the OECD PIAC and the Confemen PASEC surveys. *Correspondence*: Christian Monseur, University of Liège, Boulevard du Rectorat 5, B32, B-4000 Liege, Belgium (cmonseur@ulg.ac.be).