IMPACT OF TEST CHARACTERISTICS ON GENDER EQUITY INDICATORS IN THE ASSESSMENT OF READING COMPREHENSION

DOMINIQUE LAFONTAINE

University of Liège, Belgium

CHRISTIAN MONSEUR

University of Liège, Belgium

Australian Council for Educational Research

ABSTRACT

In this paper we discuss how apparently simple indicators such as gender differences need to be interpreted with extreme care. In particular, we consider how the assessment framework (for instance, stimulus for the assessment, types of texts, question format) and the methodology (for instance, design, definition of the target population) of international surveys may have a potential impact on the results and on the indicators. Through analysis of 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.

INTRODUCTION

Nowadays, it is current practice to build efficiency and equity indicators of educational systems based on national or international surveys. International agencies, like the OECD, the EU or the International Association for the Evaluation of Student Achievement (IEA) regularly release updated sets of indicators (see for instance *Education at a Glance*, *Key Figures on Education in the European Union*). 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 field 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 contextualized 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.

Several characteristics of the surveys have a potential impact on the results and on the indicators: (i) the assessment framework (for instance, stimulus for the assessment, types of texts, question format) and (ii) the methodological framework (for instance, design, definition of the target population).

Nowadays, 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 PISA 2003 international report (OECD, 2004) presents in a figure the gender difference on the combined reading literacy scale for the 2000 and 2003 assessments. One might also compare the gender differences between the IEA Reading literacy study and the PISA 2000 study.

Such comparisons assume that (i) the assessment framework and the survey methodology have no major or significant impact on the difference in gender performance or (ii) both assessments are fully comparable on these two aspects. If not, then an apparent increase or decrease in the achievement of some groups of students (be it a whole country or some subgroups like males and females) could, at least partially, result from variations of some influential characteristics of the framework of the respective assessments.

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 texts, 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 IEA reading Literacy and the PISA 2000 studies. In 1991, the gender gap in IEA reading comprehension among 14 years old 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 ten 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 paper will focus on parts of the assessment framework, and in particular on the question format, the reading process and the type of texts.

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; OECD, 2001, Adams & Wu, 2002), will be used. Most of the statistical analyses will be led on the PISA data, but 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

IEA Reading Literacy Study was conducted in 1991 in 31 educational systems and aimed at assessing reading comprehension among 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 comprises 89 items all included in a single booklet administered to all sample students.

The PISA 2000 study was implemented in 32 countries[1] 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).

The main aspects of the two studies are presented side by side in the following tables.

Population definition

The IEA Reading Literacy study and the OECD PISA 2000 study differ in their target population, as shown by Table 1.

[1] In this study, the results of Liechtenstein will not be included, as the sample size was only about 350 students.

Table 1: Target population

	IEA Reading Literacy (1991)	PISA 2000
Population definition	Grade attended by the majority of 14 year-olds	15 year-olds, regardless of the grade attended

Students in PISA 2000 are somewhat older, but the grade population *versus* the age population constitutes the main difference. This has no consequence 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, Lafontaine and Monseur (2004) 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. As an example, Table 2 presents the distribution of the PISA 2000 sample of the French Speaking Community of Belgium, by gender and by grade. More than 60 percent of the females are in the expected grade, *i.e.* grade 10 in this example, but less than 50 percent of the males are attending the expected grade.

Table 2: Distribution of the PISA 2000 sample of the French Speaking Community of Belgium, by gender and by grade

Grade	7	8	9	10	11	12	Total
Females	0.60	7.35	28.40	62.38	1.27	0.00	100
Males	0.17	9.43	39.43	49.95	0.89	0.13	100
Total	0.38	8.39	33.93	56.14	1.08	0.07	100

Stimulus for the reading tasks

Both studies have used continuous and non continuous texts as reading stimulus, and a variety of types of texts.

Table 3: Proportion of items for the various types of texts

	IEA Reading Literacy (1991)	PISA 2000
Continuous texts		
narrative	33 %	13 %
expository/descriptive	29 %	35 %
argumentative/injunctive	0 %	18 %
Non continuous texts	38 %	34 %

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

Reading aspects assessed

In both studies, several reading aspects or processes have been assessed.

Table 4: Proportion of items assessing the following aspects[2] as defined by PISA (OECD, 1999)[3]

	IEA Reading Literacy (1991)	PISA 2000
Retrieving or locating information (literal or paraphrase)	42 %	30 %
Interpreting (inferring, finding the main idea)	58 %	50 %
Reflecting and evaluation	0 %	20 %

[2] The PISA 2000 Initial report (OECD, 2001) defines (i) retrieving information as locating one or more information pieces of information in a text, (ii) interpreting texts as constructing meaning and drawing inferences from one or more parts of a text and (iii) reflecting and evaluation as relating a text to one's experience, knowledge and ideas. For more details, see OECD (2001).

[3] The items from IEARLS have been classified 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.

Table 5: Proportion of items by question format

	IEA Reading Literacy (1991)	PISA 2000
Multiple-choice	75 %	45 %
Open-ended short answer	22 %	11 %
Constructed open-ended	3 %	45 %

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 open-ended and the scoring relies on detailed correction procedures.

REVIEW OF THE LITERATURE

Extensive research has been dedicated to the effect of item format on achievement of males and females, including the mountain of 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 many researches have supported this general finding (DeMars, 2000 ; Mazzeo *et al*., 1991), other studies have shown that this pattern doesn't hold true in all subject areas. Mullis *et al.* (2000) have analysed gender differences by item format in IEA's Third International Mathematics and Science Study (TIMSS)[4]. There were three different types of item format:

[4] Students from 41 countries have been tested in 1994-95 in mathematics and science at 4th grade (primary school), 8th grade (middle school) and final grade of secondary school (Beaton *et al.*, 1996).

multiple-choice, short answer and extended response. There were few significant differences: almost no significant difference for either subject 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 subjects 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 and Turner (2003), 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*" (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). This last finding could explain why the interaction between gender and item format is nearly always observed in subject areas in which girls traditionally outperformed 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

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).

Furthermore, as stated by Bennett (1993), there is a "*considerable potential for interaction effects*" (p. 23); so we hypothesise that additional interaction effects may be observed and that the impact of question format may be larger according to:

1. the reading aspect (retrieve, interpret, reflect upon the text): cognitive demands for questions assessing reflection, and to a lesser extent, interpretation of the text, are higher than for locating or retrieving interpretation. Therefore, one could think that the cognitive processes assessed through multiple-choice or constructed answers should be more divergent for more cognitively demanding questions and consequently the width of the gap should be larger for the aspect "reflect" than for "interpreting the text" and for "locating information" (hypothesis 2);

2. 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

PISA 2000 released international data base contains 5 subscales for reading literacy: (i) reading retrieving information, (ii) reading interpreting, (iii) reading reflecting, (iv) reading continuous texts and (v) reading non continuous texts.

The mixed coefficients parameters multinomial logit model as described by Adams, Wilson and Wang (1997) was used to scale the PISA data and implemented by Conquest software (Wu, Adams & Wilson, 1997). This model is a generalised form of the Rasch model. For more details see Adams and Wu (2002).

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

1. the PISA 2000 initial subscales were generated according to a multidimensional model while the ten new subscales were generated according to a one dimensional model;

2. the PISA 2000 initial subscales were conditioned on all student level background variables while the new subscales were only conditioned

on the gender variable and on the school performance mean on each new subscale[5].

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 ten subscales are:

1. reading – retrieving information – multiple-choice items;
2. reading – retrieving information – open-ended items;
3. reading – interpreting – multiple-choice items;
4. reading – interpreting – open-ended items;
5. reading – reflecting – multiple-choice items;
6. reading – reflecting – open-ended items;
7. reading – Continuous Texts – multiple-choice items;
8. reading – Continuous Texts – open-ended items;
9. reading – Non Continuous Texts – multiple-choice items;
10. reading – Non Continuous Texts – open-ended items.

In other words, the ten new subscales simply decompose the 5 initial subscales by item format. Table 6 and Table 7 present, by subscale, the number of items.

[5] The conditioning also included the booklet identification for counterbalancing the booklet effect observed on the PISA 2000 data. For more information, see Adams and Wu, 2002.

Table 6: 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 / retrieving information	12 (16.7)	24 (19.3)	36
Reading / interpreting	43 (29.8)	21 (34.2)	64
Reading / reflecting	5 (13.5)	24 (15.5)	29
Total	60	69	129

Table 7: Distribution of the PISA 2000 reading items by text type and by item format

	Mutiple Choice Items	**Open-Ended Items**	**Total**
Reading / Continuous	45 (40.5)	42 (46.5)	87
Reading / Non Continuous	15 (19.5)	27 (22.5)	42
Reading / Total	60	69	129

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 ($p < 0.001$) but the independence test between item format and text type is not rejected ($p=0.09$).

The distribution of the PISA 2000 items between reading aspects and question format is not balanced, as shown by Table 6 and its associated Chi-square. This unbalanced design therefore confounded 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.

The numbers presented in brackets represent the expected number of items that would allow interpreting the differences of the gender gap between subscales as an effect of the reading aspect. The comparison between observed and expected numbers of items locates 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 poor balance, 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 hypothesis 2 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 the imbalances for the reading aspects.

RESULTS

Before presenting the results of the analyses, the difference in reading proficiency between males and females in IEARL (1991) and PISA 2000 will be briefly reviewed. Total scores and scores on the various subscales available are presented.

In IEARLS, on average, girls significantly outperform boys by 7 points ($p < 0.05$). In 13 countries out of the 31, the gender difference is not significant ($p > 0.05$). There is no difference for non continuous texts. The difference between boys and girls are respectively equal to 3 and to 18 for informative texts and for narrative texts. The difference is therefore larger for narrative texts. No subscales are available for reading aspects/processes.

Table 8 presents the standardized gender differences on the PISA 2000 subscales. As the international standard deviation was different for each subscale, the standardized difference has been preferred.

Table 8: Standardized gender difference in PISA 2000 (OECD, 2001)

Reading subscale	Standardized Gender difference
Combined scale	0.32
Retrieve information	0.23
Interpret	0.28
Reflect	0.41
Continuous texts	0.39
Non continuous texts	0.17

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. Exhibit 1 and Exhibit 2 in the Appendices present the variation in student performance on the reading / multiple-choice items scale and on the reading / open-ended items scale[6]. In all countries, the standardized gender difference is higher for the open-ended items than it is for the multiple-choice items. The median of these standardized gender differences is respectively 0.20 and 0.28 for the multiple-choice item scale and for the open-ended scale. Hypothesis 1 is thus confirmed.

However, there are large differences in the country profiles. The relative increase of the standardized difference (*i.e.* the standardized difference for multiple-choice items divided by the standardized 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) to 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 8 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.

[6] These two sets of Plausible Values were generated by R.J. Adams, from ACER.

Exhibit 3 to Exhibit 8 in the Appendices present the variation of student achievement by aspect and by item format.

Table 9: Median of gender differences per reading aspect and per item format

	Retrieving		**Interpreting**		**Reflecting**	
	MCQ	Open-Ended	MCQ	Open-Ended	MCQ	Open-Ended
Median	20	29	29	35	33	47
Median (standardized)	0.20	0.26	0.31	0.34	0.29	0.45

As shown by Table , the gender differences vary according to the item format and according to the reading aspect. As expected, the smaller difference is associated with the retrieving aspect assessed only by multiple-choice items and the highest is associated with the reflecting 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 0.20 to 0.45 standard deviations. These results confirm hypothesis 2, *i.e.* the gender gap is higher for the aspect "reflect" than for "interpreting the text" and for "retrieving information". Although at first glance Table results would suggest that item type has a stronger influence in the "reflecting" aspect than in the two other aspects, such an interpretation would be tenuous as the reading / reflecting multiple-choice scale only consists of 5 items.

Exhibit 9 to Exhibit 12 in the Appendices present the variation of student achievement by text type and by item format. Table 10 presents the median of the gender difference and its standardized equivalent by text type and by item format.

Table 10: Median of gender differences by text type and by item format

	Continuous Texts		**Non Continuous Texts**	
	MCQ	Open-Ended	MCQ	Open-Ended
Median	34	46	11.5	20.5
Median (standardized)	0.35	0.43	0.10	0.19

Table 10 results confirm hypothesis 3, *i.e.* the gender gap according to question format is higher for continuous than for non continuous texts.

Two variance analyses were performed for summarising the results. In both analyses, the dependent variable was the non standardized 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[7] was also added to both analyses for controlling all sources of variation.

Table 11 presents the decomposition of the sum of squares of the gender differences for the two analyses.

First of all, even if the country effect substantially varies between the two analyses (respectively 0.465% and 0.280% of the total sum of squares), it remains in both cases below 50 percent. In other words, the gender difference is not only a country characteristic; it also depends on test characteristics and their eventual interactions with the country variable.

The text type appears to have the larger effect on the gender differences. More than 50 percent of the sum of squares is attributable to the text type.

The item format and its associated interactions explain about 20 to 25 percent of the gender differences observed in the countries.

Table 11 : Decomposition of the sum of squares of the gender differences

Effect	% of SS	Effect	% of SS
Country (cnt)	0.465	Country (cnt)	0.280
Process (pro)	0.241	Text (tex)	0.533
Question format (ques)	0.161	Question format (que)	0.121
Cnt * Pro	0.030	Cnt * tex	0.029
Cnt * Ques	0.049	Cnt * que	0.023
Pro * Ques	0.017	Pro * que	0.006
Cnt * pro * Ques	0.036	Cnt * tex * que	0.008

[7] As shown by Table 11, the small percentage of the different interactions that involve country reflects some uniformity in the country profiles. In other word, broadly speaking, one cannot argue that some countries have higher gender differences on multiple-choice items than on open-ended items, for instance.

CONCLUSIONS

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 ”reflecting upon the text” than for “interpreting the text” and “retrieving 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 (5 multiple-choice items for the aspect ”reflect”).

The variance analysis 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).

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

DISCUSSION

Coming back to the initial question behind this study, 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 not inconsiderable.

Besides, the impact of another parameter of the framework – the reading aspects assessed - also accounts for about a quarter of the variance of the gender gap achievement. The type of texts also appears to be one of the major components of the assessment framework that affects the gender equity indicator.

Furthermore, Lafontaine and Monseur (2004) have shown that the choice of the population definition in terms of grade vs age also has a limited[8] but significant impact on the width 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 nineties and 2000.

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 clearly not 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 organizations (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). Considering the findings of this study, this new perspective which opens the way for truly comparable assessments no doubt constitutes substantial progress in monitoring education systems on reliable grounds. However, test developers should be careful to guarantee a similar balance of the various components of the reading framework in successive assessments; otherwise the validity of the trend indicator is likely to be jeopardized. This issue might partially explain why PISA 2003 used different linear transformations to anchor the 2003 student reading performance on the PISA 2000 combined reading scale (OECD, 2005).

[8] This impact is limited to education systems in which there are high rates of grade repetition.

REFERENCES

Adams, R.J., Wilson, M.R., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21,* 1-24.

Adams, R.J., & Wu, M. (Eds.) (2002). *PISA 2000 Technical Report*. Paris : OECD.

Beaton, A.E., Mullis, I.V.S., Martin, M.O., Gonzales, E.J., Kelly, D.L., & Smith, T.A. (1996). *Mathematics achievement in the middle school years : IEA's Third International Mathematics and Science Study (TIMSS)*. Chestnut Hill, MA, USA: Boston College, TIMSS International Study Center.

Bennett, R.E. (1993). On the meaning of constructed response. R.E. Benett and W.C. Ward (Eds). *Construction versus choice in cognitive measurement. Issues in constructed response, performance testing, and portfolio assessmen*t. Hillsdale : Lawrence Erlbaum Associates. &-29.

DeMars, C. E. (2000). Test stakes and Item Format Interaction. *Applied Measurement in Education,* Vol. 13*, 1*, 55-77.

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

Kirsch, I., de Jong, J., Lafontaine, D., McQueen, J., Mendelovits, J., & Monseur, C. (2003). *La lecture, moteur de changement. Performances et engagement. Résultats de PISA 2000*. Paris : Ocdé.

Lafontaine, D., & Monseur, C. (2003). Influence des caractéristiques de l'évaluation sur les indicateurs d'équité. *Communication au 16e colloque international de l'Admée-Europe du 4 au 6 septembre 2003 à l'Université de Liège.*

Mazzeo, J., Schmitt, A.P., & Bleistein, C.A. (1991, April). *Do women peform 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.

Monseur, C., & Demeuse, M. (2004). Quelques réflexions méthodologiques à propos des enquêtes internationales dans le domaine de l'éducation.

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.* Chesnut Hill (Massachussetts, USA): TIMSS International Study Center, Boston College.

Oecd (1999). *Mesurer les connaissances et les compétences des élèves. Un nouveau cadre d'évaluation. PISA.* Paris : Ocdé.

Oecd (2001). *Connaissances et compétences: des atouts pour la vie. Premiers résultats de PISA 2000. Enseignement et compétences*. Paris : Ocdé.

Oecd (2004). *Learning for Tomorrow's World: First results from PISA 2003*. Paris : Oecd.

Oecd (2005). *PISA 2003 Technical Report*. 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 AERA.

Traub, R.E., & MacRury, K. (1990). Antwort-auswahl –vs freie-antwort-aufgaben bei lernerfolgs-test. In K. Ingenkamp & R.S. Jäger 5 Eds). *Tests und trends 8: Jahrbuch der pädagogishen diagnostik* (pp.128-159). Weinheim, Germany: Beltz Verlag.

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

APPENDICES

Exhibit 1: Variation in student performance on the reading / multiple-choice items scale (OECD, PISA 2000 database)

	Mean	SE	STD	SE	Gender difference	SE	Standardized difference
AUS	533	(2.9)	112	(2.1)	26	(5.6)	0.24
AUT	503	(2.1)	97	(1.6)	13	(4.7)	0.14
BEL	517	(3.1)	115	(2.3)	25	(5.6)	0.21
BRA	392	(2.6)	88	(1.5)	8	(4.3)	0.09
CAN	531	(1.4)	101	(0.8)	20	(1.8)	0.20
CHE	501	(3.2)	104	(1.6)	18	(4.7)	0.17
CZE	503	(1.9)	93	(1.3)	21	(3.8)	0.23
DEU	491	(2.4)	108	(1.7)	21	(5.0)	0.20
DNK	495	(2.4)	102	(1.4)	13	(4.1)	0.13
ESP	500	(2.2)	91	(1.3)	15	(3.5)	0.17
FIN	548	(2.1)	97	(1.3)	35	(3.2)	0.36
FRA	506	(2.5)	103	(1.8)	19	(3.7)	0.18
GBR	512	(2.5)	106	(1.4)	16	(3.9)	0.15
GRC	471	(3.6)	93	(1.8)	23	(4.2)	0.25
HUN	483	(3.0)	94	(1.4)	19	(4.9)	0.20
IRL	526	(2.8)	100	(1.7)	18	(4.6)	0.18
ISL	512	(2.1)	98	(1.5)	22	(3.5)	0.23
ITA	489	(2.3)	93	(1.5)	24	(5.3)	0.26
JPN	520	(3.8)	90	(1.7)	20	(5.0)	0.22
KOR	523	(1.9)	77	(1.1)	6	(5.4)	0.07
LUX	451	(1.9)	108	(1.4)	20	(4.4)	0.18
LVA	451	(4.1)	103	(2.0)	36	(4.1)	0.35
MEX	414	(2.4)	81	(1.4)	9	(4.0)	0.11
NLD	545	(3.1)	103	(2.8)	22	(6.6)	0.21
NOR	506	(2.7)	110	(1.6)	28	(4.3)	0.25
NZL	533	(2.7)	116	(2.1)	32	(6.0)	0.28
POL	475	(3.4)	100	(2.2)	24	(6.0)	0.24
PRT	479	(3.6)	96	(1.5)	18	(3.9)	0.19
RUS	451	(3.4)	98	(1.4)	25	(2.9)	0.25
SWE	527	(2.6)	105	(1.6)	24	(3.0)	0.23
USA	508	(6.1)	111	(2.0)	19	(4.5)	0.17

Exhibit 2: Variation in student performance on the reading / open-ended items scale (OECD, PISA 2000 database)

	Mean	SE	STD	SE	Gender difference	SE	Standardized difference
AUS	528	(3.2)	104	(1.8)	30	(5.5)	0.29
AUT	508	(2.2)	96	(1.5)	25	(4.8)	0.26
BEL	509	(3.0)	108	(2.0)	31	(5.7)	0.28
BRA	391	(3.0)	97	(2.0)	16	(4.3)	0.16
CAN	535	(1.4)	96	(1.0)	30	(1.9)	0.31
CHE	492	(3.9)	108	(2.0)	29	(4.4)	0.27
CZE	492	(2.2)	96	(1.5)	33	(4.2)	0.35
DEU	488	(2.5)	112	(2.1)	31	(4.6)	0.27
DNK	499	(2.4)	101	(1.9)	26	(3.8)	0.25
ESP	490	(2.4)	89	(1.4)	24	(3.7)	0.27
FIN	549	(2.0)	89	(1.3)	45	(2.8)	0.51
FRA	505	(2.5)	94	(1.6)	27	(3.6)	0.28
GBR	530	(2.5)	101	(1.6)	24	(4.4)	0.23
GRC	475	(4.8)	106	(2.5)	38	(5.1)	0.35
HUN	483	(3.7)	96	(1.9)	27	(5.9)	0.28
IRL	529	(3.1)	96	(1.8)	27	(4.2)	0.28
ISL	506	(1.8)	98	(1.6)	40	(3.2)	0.41
ITA	489	(2.6)	94	(2.0)	33	(6.4)	0.35
JPN	527	(4.7)	92	(2.7)	28	(6.1)	0.30
KOR	527	(1.9)	73	(1.3)	11	(4.8)	0.15
LUX	440	(1.8)	112	(1.6)	29	(4.3)	0.26
LVA	463	(5.2)	112	(2.2)	53	(4.6)	0.47
MEX	428	(3.1)	96	(1.8)	20	(4.5)	0.21
NLD	534	(3.0)	86	(2.2)	25	(5.1)	0.29
NOR	507	(2.6)	107	(1.7)	41	(4.0)	0.39
NZL	529	(2.6)	109	(1.8)	45	(6.5)	0.41
POL	478	(4.5)	107	(2.8)	35	(7.0)	0.33
PRT	466	(4.3)	105	(2.1)	23	(4.0)	0.22
RUS	468	(3.6)	95	(1.3)	35	(2.9)	0.37
SWE	513	(2.2)	95	(1.3)	34	(2.8)	0.36
USA	501	(6.5)	108	(2.5)	30	(4.4)	0.28

Exhibit 3: Variation in student performance on the reading / retrieving information – multiple-choice items scale (OECD, PISA 2000 database)

	Mean	SE	STD	SE	Gender difference	SE	Standardized difference
AUS	542	(4.0)	112	(2.4)	35	(7.0)	0.31
AUT	505	(2.5)	96	(1.8)	3	(6.4)	0.03
BEL	511	(4.4)	124	(3.4)	19	(6.5)	0.15
BRA	374	(3.2)	82	(2.0)	1	(3.5)	0.01
CAN	523	(1.8)	101	(1.0)	23	(2.4)	0.23
CHE	508	(4.3)	113	(2.8)	18	(4.8)	0.16
CZE	480	(3.0)	90	(3.2)	20	(4.8)	0.22
DEU	493	(3.4)	106	(2.8)	12	(4.9)	0.11
DNK	498	(3.0)	107	(2.1)	8	(5.6)	0.07
ESP	492	(3.1)	86	(1.4)	13	(3.9)	0.15
FIN	548	(2.4)	89	(1.5)	37	(2.7)	0.42
FRA	506	(3.2)	101	(2.3)	15	(4.1)	0.15
GBR	525	(2.9)	104	(2.4)	17	(4.1)	0.16
GRC	448	(4.7)	95	(3.1)	24	(4.9)	0.25
HUN	485	(4.8)	109	(2.9)	23	(6.3)	0.21
IRL	523	(3.3)	90	(1.9)	18	(5.5)	0.20
ISL	499	(1.9)	93	(1.9)	20	(3.9)	0.22
ITA	491	(3.4)	103	(2.1)	20	(8.2)	0.19
JPN	532	(5.4)	93	(3.1)	24	(6.4)	0.26
KOR	533	(2.8)	83	(1.7)	2	(7.9)	0.02
LUX	446	(2.1)	111	(1.7)	11	(4.0)	0.10
LVA	444	(5.1)	96	(2.7)	28	(4.7)	0.29
MEX	386	(3.4)	80	(2.1)	9	(4.7)	0.11
NLD	545	(4.8)	108	(3.6)	25	(6.8)	0.23
NOR	514	(3.5)	113	(1.8)	21	(4.4)	0.19
NZL	541	(3.4)	119	(2.1)	42	(6.8)	0.35
POL	471	(4.7)	96	(3.2)	21	(6.7)	0.22
PRT	463	(4.6)	97	(2.2)	11	(3.5)	0.11
RUS	447	(4.9)	99	(2.7)	28	(3.6)	0.28
SWE	521	(2.7)	108	(1.5)	29	(3.6)	0.27
USA	497	(6.1)	105	(2.8)	24	(4.3)	0.23

Exhibit 4: Variation in student performance on the reading / retrieving information – open-ended items scale (OECD, PISA 2000 database)

	Mean	SE	STD	SE	Gender difference	SE	Standardized difference
AUS	530	(3.8)	113	(1.9)	29	(5.9)	0.26
AUT	501	(2.9)	101	(1.7)	22	(6.0)	0.22
BEL	518	(4.0)	125	(3.0)	29	(6.4)	0.23
BRA	355	(3.8)	111	(2.2)	18	(4.7)	0.16
CAN	531	(1.8)	103	(1.2)	28	(2.0)	0.27
CHE	500	(4.7)	117	(2.5)	26	(5.0)	0.22
CZE	478	(3.2)	119	(2.4)	34	(5.7)	0.29
DEU	477	(3.0)	123	(2.8)	38	(5.4)	0.31
DNK	498	(2.9)	109	(2.3)	20	(4.1)	0.18
ESP	482	(3.2)	94	(1.5)	20	(3.7)	0.21
FIN	565	(3.4)	111	(3.4)	50	(3.6)	0.45
FRA	519	(3.2)	105	(2.2)	28	(3.9)	0.27
GBR	523	(3.2)	107	(2.0)	25	(5.4)	0.23
GRC	450	(5.5)	116	(3.4)	32	(5.7)	0.28
HUN	466	(4.5)	111	(2.1)	25	(6.7)	0.23
IRL	523	(3.7)	104	(1.6)	24	(4.9)	0.23
ISL	502	(2.2)	113	(1.9)	41	(3.4)	0.36
ITA	487	(3.5)	108	(3.1)	34	(7.7)	0.31
JPN	525	(5.5)	105	(3.2)	31	(7.4)	0.30
KOR	527	(2.8)	86	(1.9)	9	(6.6)	0.10
LUX	432	(1.9)	116	(2.0)	29	(4.8)	0.25
LVA	455	(5.8)	134	(2.7)	53	(6.0)	0.40
MEX	414	(4.7)	116	(2.9)	16	(6.1)	0.14
NLD	548	(4.0)	95	(2.9)	19	(6.5)	0.20
NOR	500	(2.9)	112	(2.0)	38	(4.1)	0.34
NZL	531	(3.4)	118	(2.3)	39	(7.2)	0.33
POL	470	(5.5)	124	(4.1)	31	(8.8)	0.25
PRT	455	(5.1)	112	(2.4)	21	(4.2)	0.19
RUS	445	(5.2)	116	(2.3)	37	(4.6)	0.32
SWE	514	(2.5)	101	(1.7)	30	(3.2)	0.30
USA	495	(7.1)	115	(3.4)	31	(5.2)	0.27

Exhibit 5: Variation in student performance on the reading / interpreting texts – multiple-choice items scale (OECD, PISA 2000 database)

	Mean	SE	STD	SE	Gender difference	SE	Standardized difference
AUS	530	(3.8)	109	(1.7)	37	(6.1)	0.34
AUT	502	(2.7)	94	(2.0)	22	(5.5)	0.23
BEL	511	(3.8)	109	(2.8)	29	(6.3)	0.27
BRA	398	(3.1)	84	(1.5)	14	(4.1)	0.17
CAN	531	(1.7)	98	(0.9)	30	(2.0)	0.31
CHE	501	(4.0)	103	(2.4)	25	(4.3)	0.24
CZE	502	(2.5)	89	(2.1)	30	(4.0)	0.34
DEU	483	(2.9)	108	(2.3)	33	(5.0)	0.31
DNK	501	(2.5)	100	(1.6)	19	(3.4)	0.19
ESP	498	(2.7)	85	(1.2)	23	(2.8)	0.27
FIN	548	(2.8)	97	(3.0)	48	(3.7)	0.49
FRA	505	(3.1)	97	(1.9)	27	(3.6)	0.28
GBR	511	(3.0)	102	(2.0)	22	(4.8)	0.22
GRC	473	(4.2)	89	(2.2)	33	(4.8)	0.37
HUN	481	(3.9)	91	(2.4)	29	(5.2)	0.32
IRL	526	(3.3)	97	(1.4)	26	(5.4)	0.27
ISL	512	(1.6)	92	(1.9)	33	(3.3)	0.36
ITA	486	(2.6)	84	(2.0)	33	(5.7)	0.39
JPN	519	(5.0)	88	(2.4)	28	(6.4)	0.32
KOR	525	(2.8)	72	(1.5)	10	(5.9)	0.14
LUX	451	(1.8)	105	(1.7)	29	(4.3)	0.28
LVA	455	(4.7)	100	(2.6)	49	(4.5)	0.49
MEX	415	(2.9)	76	(1.8)	15	(4.0)	0.20
NLD	538	(3.9)	94	(2.9)	27	(6.2)	0.29
NOR	509	(3.0)	107	(1.8)	35	(4.1)	0.33
NZL	530	(3.0)	112	(2.0)	41	(6.3)	0.37
POL	474	(4.4)	94	(2.7)	31	(6.5)	0.33
PRT	476	(4.3)	90	(2.0)	27	(3.4)	0.30
RUS	457	(4.1)	94	(2.1)	31	(3.4)	0.33
SWE	530	(2.4)	101	(1.6)	35	(4.0)	0.35
USA	509	(7.6)	109	(2.6)	26	(4.9)	0.24

Exhibit 6: Variation in student performance on the reading / interpreting texts – open-ended items scale (OECD, PISA 2000 database)

	Mean	SE	STD	SE	Gender difference	SE	Standardized difference
AUS	527	(4.1)	111	(2.4)	38	(5.7)	0.34
AUT	516	(3.0)	100	(1.8)	25	(5.7)	0.25
BEL	517	(3.3)	109	(2.1)	36	(6.2)	0.33
BRA	398	(3.3)	92	(2.3)	16	(4.2)	0.17
CAN	534	(1.7)	98	(1.1)	35	(2.2)	0.36
CHE	495	(4.6)	107	(2.5)	30	(4.5)	0.28
CZE	507	(2.7)	102	(2.4)	40	(5.1)	0.39
DEU	499	(3.2)	118	(2.4)	40	(5.1)	0.34
DNK	492	(2.4)	106	(1.8)	30	(4.3)	0.28
ESP	484	(2.6)	85	(1.2)	23	(2.8)	0.27
FIN	572	(2.6)	96	(1.9)	57	(3.8)	0.59
FRA	512	(3.1)	93	(1.9)	32	(3.6)	0.34
GBR	524	(2.8)	105	(2.3)	28	(4.9)	0.27
GRC	480	(5.1)	100	(3.0)	38	(5.2)	0.38
HUN	484	(4.3)	96	(2.7)	31	(6.1)	0.32
IRL	533	(4.3)	104	(2.3)	32	(5.7)	0.31
ISL	520	(2.3)	109	(1.9)	53	(4.1)	0.49
ITA	495	(3.2)	94	(2.8)	44	(7.0)	0.47
JPN	528	(5.5)	88	(2.9)	25	(7.0)	0.28
KOR	533	(2.6)	72	(1.6)	12	(5.6)	0.17
LUX	442	(2.2)	119	(2.0)	35	(5.3)	0.29
LVA	472	(5.2)	98	(2.1)	56	(5.1)	0.57
MEX	426	(3.3)	88	(1.8)	22	(4.3)	0.25
NLD	533	(3.7)	95	(2.7)	39	(6.3)	0.41
NOR	505	(3.1)	110	(1.7)	51	(3.8)	0.46
NZL	527	(3.7)	116	(1.9)	52	(6.9)	0.45
POL	492	(4.8)	106	(3.4)	41	(7.7)	0.39
PRT	469	(4.8)	103	(2.5)	26	(3.7)	0.25
RUS	487	(3.7)	90	(1.9)	37	(2.5)	0.41
SWE	513	(2.8)	102	(1.9)	39	(3.4)	0.38
USA	498	(7.5)	110	(2.8)	33	(4.6)	0.30

Exhibit 7: Variation in student performance on the reading /reflection and evaluation – multiple-choice items scale (OECD, PISA 2000 database)

	Mean	SE	STD	SE	Gender difference	SE	Standardized difference
AUS	534	(5.0)	132	(2.2)	54	(8.1)	0.41
AUT	503	(4.2)	115	(2.4)	33	(6.4)	0.29
BEL	524	(6.6)	150	(5.3)	34	(8.5)	0.23
BRA	369	(3.0)	73	(1.8)	19	(3.5)	0.26
CAN	534	(2.3)	113	(1.5)	33	(2.5)	0.29
CHE	504	(5.1)	128	(2.7)	26	(5.6)	0.20
CZE	501	(4.0)	131	(2.0)	41	(6.2)	0.31
DEU	478	(4.5)	136	(3.9)	34	(6.0)	0.25
DNK	472	(3.1)	118	(1.9)	22	(4.6)	0.19
ESP	539	(4.3)	121	(2.4)	27	(5.1)	0.22
FIN	505	(2.5)	110	(2.2)	49	(4.6)	0.45
FRA	520	(4.3)	125	(2.6)	31	(5.0)	0.25
GBR	524	(3.7)	120	(1.9)	30	(5.4)	0.25
GRC	465	(6.3)	109	(3.2)	33	(5.6)	0.30
HUN	491	(4.7)	121	(3.3)	35	(7.1)	0.29
IRL	531	(4.3)	114	(2.3)	36	(7.9)	0.32
ISL	483	(1.8)	88	(1.5)	32	(3.7)	0.36
ITA	506	(4.5)	108	(3.2)	33	(9.3)	0.31
JPN	552	(6.8)	132	(3.7)	36	(8.9)	0.27
KOR	506	(3.5)	88	(1.4)	14	(6.8)	0.16
LUX	439	(2.1)	125	(2.6)	29	(5.8)	0.23
LVA	432	(5.6)	104	(2.0)	36	(5.9)	0.35
MEX	405	(4.3)	87	(2.5)	11	(4.5)	0.13
NLD	567	(5.5)	110	(4.0)	23	(8.7)	0.21
NOR	476	(3.5)	121	(1.9)	49	(4.8)	0.40
NZL	528	(4.5)	131	(2.2)	44	(7.8)	0.34
POL	464	(5.6)	118	(3.4)	36	(8.6)	0.31
PRT	486	(4.8)	106	(2.2)	19	(5.1)	0.18
RUS	422	(4.1)	100	(2.2)	33	(3.3)	0.33
SWE	522	(2.9)	116	(2.1)	44	(5.0)	0.38
USA	514	(7.3)	128	(3.3)	29	(6.0)	0.23

Exhibit 8: Variation in student performance on the reading /reflection and evaluation – open-ended items scale (OECD, PISA 2000 database)

	Mean	SE	STD	SE	Gender difference	SE	Standardized difference
AUS	527	(3.8)	99	(2.3)	41	(6.1)	0.41
AUT	513	(3.1)	99	(2.4)	41	(5.6)	0.41
BEL	496	(3.7)	112	(2.6)	47	(6.3)	0.42
BRA	423	(3.4)	97	(2.4)	28	(4.7)	0.29
CAN	543	(1.7)	94	(1.0)	47	(2.1)	0.50
CHE	488	(4.8)	111	(2.6)	48	(4.6)	0.43
CZE	482	(3.0)	108	(3.1)	57	(5.1)	0.53
DEU	476	(4.0)	143	(5.2)	58	(7.0)	0.41
DNK	510	(2.6)	98	(2.3)	46	(3.6)	0.47
ESP	506	(2.8)	88	(1.3)	40	(3.4)	0.45
FIN	531	(2.5)	86	(3.3)	64	(2.3)	0.74
FRA	496	(3.0)	92	(1.8)	40	(3.8)	0.43
GBR	543	(2.6)	93	(1.5)	34	(4.2)	0.37
GRC	498	(6.1)	120	(3.5)	57	(6.7)	0.48
HUN	478	(4.5)	103	(2.8)	44	(6.7)	0.43
IRL	536	(2.9)	83	(1.3)	37	(5.4)	0.45
ISL	505	(1.6)	90	(1.5)	57	(3.8)	0.63
ITA	479	(3.6)	102	(2.6)	48	(7.8)	0.47
JPN	531	(5.2)	98	(2.7)	43	(6.7)	0.44
KOR	529	(2.8)	75	(1.9)	30	(5.5)	0.40
LUX	440	(2.3)	112	(1.6)	49	(3.7)	0.44
LVA	463	(5.8)	118	(2.4)	75	(5.1)	0.64
MEX	450	(4.0)	117	(2.3)	40	(6.5)	0.34
NLD	525	(3.4)	79	(2.2)	36	(5.0)	0.46
NOR	514	(3.2)	104	(1.8)	60	(3.9)	0.58
NZL	531	(2.9)	103	(1.9)	58	(6.6)	0.56
POL	479	(4.8)	110	(3.2)	54	(7.8)	0.49
PRT	477	(4.8)	102	(2.2)	40	(3.6)	0.39
RUS	462	(4.0)	96	(1.8)	48	(3.1)	0.50
SWE	513	(2.6)	92	(1.2)	52	(3.2)	0.57
USA	503	(7.4)	103	(2.7)	39	(5.1)	0.38

Exhibit 9: Variation in student performance on the reading / continuous texts – multiple-choice items scale (OECD, PISA 2000 database)

	Mean	SE	STD	SE	Gender difference	SE	Standardized difference
AUS	531	(3.9)	111	(2.0)	43	(6.4)	0.39
AUT	507	(3.1)	97	(2.2)	27	(5.8)	0.28
BEL	511	(4.3)	115	(3.9)	34	(6.8)	0.30
BRA	403	(3.4)	88	(2.0)	15	(4.7)	0.17
CAN	532	(2.0)	100	(1.0)	35	(2.0)	0.35
CHE	503	(4.3)	107	(2.5)	32	(4.4)	0.30
CZE	491	(2.6)	92	(2.4)	35	(4.2)	0.38
DEU	484	(3.1)	111	(2.2)	40	(5.6)	0.36
DNK	502	(3.3)	105	(2.0)	26	(3.7)	0.25
ESP	498	(3.0)	85	(1.2)	25	(3.3)	0.29
FIN	540	(3.3)	97	(3.7)	55	(3.0)	0.57
FRA	504	(3.2)	100	(2.0)	31	(3.8)	0.31
GBR	513	(3.1)	104	(1.8)	27	(4.7)	0.26
GRC	473	(4.9)	96	(2.8)	38	(5.2)	0.40
HUN	481	(3.9)	93	(2.4)	32	(5.5)	0.34
IRL	532	(3.5)	99	(1.5)	28	(5.2)	0.28
ISL	511	(1.7)	97	(1.5)	34	(3.3)	0.35
ITA	494	(2.8)	91	(2.2)	38	(6.5)	0.42
JPN	521	(4.9)	86	(2.5)	31	(6.1)	0.36
KOR	526	(2.9)	76	(1.8)	12	(6.9)	0.16
LUX	454	(1.8)	108	(1.4)	34	(4.3)	0.31
LVA	453	(4.8)	101	(2.2)	47	(4.6)	0.47
MEX	417	(3.2)	81	(2.1)	18	(4.3)	0.22
NLD	538	(4.2)	100	(3.0)	35	(7.3)	0.35
NOR	511	(3.4)	109	(1.6)	44	(4.0)	0.40
NZL	532	(3.7)	114	(2.4)	47	(6.6)	0.41
POL	472	(4.5)	98	(2.9)	38	(6.4)	0.39
PRT	478	(4.5)	94	(2.0)	27	(3.4)	0.29
RUS	452	(4.1)	96	(2.2)	36	(3.4)	0.38
SWE	535	(2.7)	106	(1.7)	43	(3.5)	0.41
USA	508	(7.4)	109	(2.8)	30	(4.7)	0.28

Exhibit 10: Variation in student performance on the reading / continuous texts – open-ended items scale (OECD, PISA 2000 database)

	Mean	SE	STD	SE	Gender difference	SE	Standardized difference
AUS	523	(3.8)	110	(2.0)	43	(6.1)	0.39
AUT	513	(2.9)	100	(1.8)	39	(5.5)	0.39
BEL	500	(4.2)	113	(3.7)	45	(6.3)	0.40
BRA	411	(3.7)	98	(2.2)	29	(4.0)	0.30
CAN	539	(1.8)	101	(1.4)	47	(2.0)	0.47
CHE	490	(4.9)	115	(2.7)	49	(4.6)	0.43
CZE	487	(2.8)	107	(2.3)	56	(4.6)	0.52
DEU	483	(3.2)	131	(2.7)	56	(6.1)	0.43
DNK	496	(2.9)	110	(2.5)	45	(3.8)	0.41
ESP	492	(2.8)	88	(1.3)	38	(3.3)	0.43
FIN	551	(3.0)	96	(3.9)	68	(2.9)	0.71
FRA	498	(3.2)	98	(2.1)	41	(3.9)	0.42
GBR	529	(2.9)	106	(1.9)	37	(5.0)	0.35
GRC	487	(5.7)	115	(3.5)	57	(6.2)	0.50
HUN	481	(4.2)	101	(2.2)	44	(6.4)	0.44
IRL	528	(3.6)	98	(2.3)	42	(5.2)	0.43
ISL	508	(2.0)	102	(1.6)	58	(3.7)	0.57
ITA	489	(3.6)	101	(3.4)	52	(7.8)	0.51
JPN	531	(5.6)	99	(3.2)	42	(7.3)	0.42
KOR	535	(2.7)	72	(2.3)	25	(5.7)	0.35
LUX	433	(1.7)	120	(1.4)	49	(4.5)	0.41
LVA	469	(6.1)	120	(2.9)	72	(5.3)	0.60
MEX	445	(3.5)	104	(2.0)	34	(4.7)	0.33
NLD	523	(3.6)	90	(2.8)	41	(5.9)	0.46
NOR	506	(3.3)	110	(2.1)	64	(4.4)	0.58
NZL	526	(3.5)	117	(2.5)	61	(7.3)	0.52
POL	488	(5.1)	114	(3.7)	54	(8.0)	0.47
PRT	472	(5.2)	110	(2.5)	40	(4.3)	0.36
RUS	476	(4.3)	100	(2.2)	51	(3.0)	0.51
SWE	506	(2.6)	96	(1.6)	52	(3.4)	0.54
USA	499	(8.3)	113	(4.1)	43	(5.5)	0.38

Exhibit 11: Variation in student performance on the reading / non continuous texts – multiple-choice items scale (OECD, PISA 2000 database)

	Mean	SE	STD	SE	Gender difference	SE	Standardized difference
AUS	557	(5.2)	125	(2.3)	26	(7.3)	0.21
AUT	502	(3.2)	107	(2.2)	-6	(6.6)	-0.06
BEL	520	(4.9)	127	(3.8)	15	(7.3)	0.12
BRA	367	(4.1)	101	(2.3)	-1	(4.2)	-0.01
CAN	538	(2.3)	109	(1.2)	12	(2.7)	0.11
CHE	512	(5.0)	124	(3.1)	5	(5.3)	0.04
CZE	517	(3.8)	120	(3.6)	12	(6.3)	0.10
DEU	496	(3.2)	117	(2.8)	4	(5.5)	0.03
DNK	503	(3.4)	124	(2.9)	-1	(4.5)	-0.01
ESP	502	(3.7)	103	(1.5)	12	(4.0)	0.12
FIN	560	(3.4)	112	(1.9)	25	(4.1)	0.22
FRA	527	(3.9)	117	(2.6)	6	(4.7)	0.05
GBR	533	(3.4)	113	(2.4)	9	(5.7)	0.08
GRC	450	(4.3)	90	(2.2)	13	(4.5)	0.14
HUN	487	(5.5)	125	(3.1)	18	(7.6)	0.14
IRL	526	(4.5)	109	(2.2)	15	(6.7)	0.14
ISL	512	(1.5)	102	(1.8)	21	(3.9)	0.21
ITA	487	(4.0)	106	(2.8)	11	(8.9)	0.10
JPN	524	(6.7)	116	(3.5)	18	(8.7)	0.16
KOR	516	(3.1)	82	(2.0)	-5	(7.8)	-0.06
LUX	446	(1.8)	118	(2.1)	3	(4.7)	0.03
LVA	446	(6.1)	115	(3.2)	32	(5.6)	0.28
MEX	395	(3.4)	86	(2.2)	-2	(4.4)	-0.02
NLD	552	(5.2)	105	(3.3)	6	(7.1)	0.06
NOR	513	(4.3)	127	(2.4)	11	(5.4)	0.09
NZL	553	(4.0)	133	(2.3)	38	(7.7)	0.29
POL	476	(5.8)	110	(2.8)	1	(8.0)	0.01
PRT	460	(5.0)	103	(2.9)	10	(3.7)	0.10
RUS	454	(5.6)	118	(3.0)	22	(3.7)	0.19
SWE	523	(3.8)	113	(1.9)	19	(4.0)	0.17
USA	515	(8.8)	121	(3.6)	12	(5.1)	0.10

Exhibit 12: Variation in student performance on the reading / non continuous texts – open-ended items scale (OECD, PISA 2000 database)

	Mean	SE	STD	SE	Gender difference	SE	Standardized difference
AUS	537	(4.2)	103	(2.3)	21	(6.4)	0.20
AUT	513	(2.9)	100	(1.9)	12	(6.2)	0.12
BEL	521	(3.5)	114	(2.5)	23	(6.5)	0.20
BRA	371	(4.0)	98	(2.3)	7	(5.3)	0.07
CAN	539	(1.8)	97	(1.2)	21	(2.2)	0.22
CHE	497	(4.9)	109	(2.8)	11	(4.6)	0.10
CZE	498	(3.1)	114	(3.1)	27	(6.1)	0.24
DEU	486	(3.0)	117	(2.5)	23	(5.4)	0.20
DNK	504	(3.4)	109	(3.0)	11	(4.5)	0.10
ESP	490	(3.4)	98	(1.8)	13	(4.5)	0.13
FIN	556	(3.3)	96	(3.9)	38	(3.4)	0.40
FRA	524	(3.1)	93	(2.0)	17	(3.8)	0.18
GBR	538	(2.9)	96	(1.8)	14	(5.4)	0.15
GRC	459	(5.6)	114	(3.5)	22	(6.1)	0.19
HUN	478	(4.7)	106	(2.2)	17	(6.5)	0.16
IRL	534	(3.6)	99	(1.6)	17	(5.3)	0.17
ISL	506	(2.0)	104	(1.7)	35	(3.7)	0.34
ITA	479	(3.8)	103	(3.1)	24	(8.6)	0.23
JPN	522	(5.6)	95	(2.9)	22	(7.3)	0.23
KOR	512	(2.9)	83	(1.9)	7	(6.6)	0.08
LUX	446	(1.7)	115	(1.3)	19	(3.9)	0.17
LVA	449	(5.4)	116	(3.3)	46	(5.8)	0.40
MEX	400	(4.8)	112	(2.6)	13	(5.9)	0.12
NLD	544	(3.6)	82	(2.8)	15	(5.5)	0.18
NOR	511	(3.6)	109	(2.1)	27	(3.8)	0.25
NZL	539	(3.2)	106	(1.9)	33	(7.2)	0.31
POL	473	(5.3)	111	(4.0)	21	(8.0)	0.19
PRT	461	(4.8)	104	(2.5)	15	(3.7)	0.14
RUS	454	(4.9)	100	(2.4)	26	(3.5)	0.26
SWE	522	(3.0)	100	(1.4)	20	(3.4)	0.20
USA	506	(7.2)	107	(3.4)	23	(4.8)	0.21