Large-scale Assessments in Education
a SpringerOpen Journal

RESEARCH Open Access

Gender differences in variability and extreme scores in an international context

Ariane Baye* and Christian Monseur

*Correspondence: ariane.baye@ulg.ac.be
Department of Education and Training, Faculty of Psychology, Speech Therapy, and Education, University of Liège, 2, Place des Orateurs, Quartier Agora, 4000 Liège, Belgium

Abstract

This study examines gender differences in the variability of student performance in reading, mathematics and science. Twelve databases from IEA and PISA were used to analyze gender differences within an international perspective from 1995 to 2015. Effect sizes and variance ratios were computed. The main results are as follows. (1) Gender differences vary by content area, students' educational levels, and students' proficiency levels. The gender differences at the extreme tails of the distribution are often more substantial than the gender differences at the mean. (2) Exploring the extreme tails of the distributions shows that the situation of the weakest males in reading is a real matter of concern. In mathematics and science, males are more frequently among the highest performing students. (3) The "greater male variability hypothesis" is confirmed.

Keywords: Gender, Effect size, PISA, PIRLS, TIMSS

The women's success paradox

Gender equity in educational outcomes is considered to be one of the major equity concerns in democratic societies (Marks 2008; Wagemaker 1996; Willingham and Cole 1997). Often associated with reduced career opportunities, gender differences in achievement are high-stake issues. One of the six key educational goals of UNESCO was to achieve gender equality in educational opportunities by 2015 (UNESCO 2011).

From an educational point of view, paying attention to gender differences in developed countries, and in particular to the situation of girls, may seem outdated. In most countries, there has been a substantial rise in gender equality over the last decades (Barro and Lee 2001). Girls have gained much with the extension of compulsory education (Baudelot and Estabelet 2007); their grade retention rate is lower than boys', they have on average better results in school, and are more likely to enter higher education (Buchmann and DiPrete 2006; OECD 2014a).

Meta-analyses on gender differences demonstrate improvements in girls' results over time. Focusing on mathematics U.S. math data (from 5 year-olds to adulthood), Hyde et al. (1990, p. 139) concluded that the "magnitude of the gender difference has declined over the years" (for studies published in 1973 or earlier Cohen *d* was 0.31, whereas it was 0.14 for studies published in 1974 or later). In international and national surveys, Lietz observed the opposite pattern for reading at the secondary level. [...] "Since 1992 girls outperformed boys to a considerably greater extent when compared with studies up to and including 1991" (Lietz 2006a, p. 139). This apparent contradiction in the evolution of gender differences in mathematics and in reading might simply reflect an improvement in females' performance, and/or a decline in that of males. As males have usually outperformed females in mathematics, an improvement in females' performance would reduce the gender gap. Conversely, as females tend to outperform males in reading, an improvement in females' performance would increase the gender gap.

Yet the female success is paradoxical (Marry 2003): some do not have access to the most prestigious higher education fields and their diplomas are less effective in labour markets (Jacobs, 1996, Dubet 2010). According to Dubet (2010), two hypotheses can be formulated to explain this paradox: the institutional hypothesis assumes that schools and teachers are still stuck in a male-dominated model where the division of roles is gendered. Students are differently oriented to the studies and professions depending on their sex, according to gender norms. The school thus reproduces societal cleavages. Another hypothesis is that girls anticipate more their family life and consciously make the choice of studies and careers that allow them to manage family and professional life.

We argue that a third hypothesis has been neglected: the female success story is viewed from a "mean" perspective (i.e. based on gender equality, on average). If the most gifted boys still outperform the most gifted girls, the low rate of female students in mathematics and science fields in higher education, as the subsequent underrepresentation of women in the related occupations, is not so paradoxical.

Moreover, following Hill et al.'s (2008) work, we argue that Cohen's (1977) rule of thumb has led to misrepresent the actual gender differences, which have to be interpreted according to empirical benchmarks. This paper contributes to provide empirical benchmarks to evaluate gender differences: effect sizes on the mean and at the extreme tails of the achievement distributions have been computed on the ten latest international large-scale assessments.

Background

International comparative surveys of student achievement such as those carried out by the International Association for the Evaluation of Educational Achievement (IEA) or by the Organisation for Economic Co-operation and Development (OECD) have for a long time been useful tools for estimating the magnitude of gender gaps, in particular for educational outcomes. Those organisations assess large and representative national samples of students (based on grade and/or age criteria), which avoid the issue of unrepresentative samples due to "the great deal of self-selection", on the basis of which generalisations have been made about gender differences (Nowell and Hedges 1998).

The popular generalisation about mean gender differences in reading, mathematics and science may be summarised using Johnson's work (1996), based on two international and six national [US] assessment programmes. Johnson's conclusions highlight a gender gap in favour of girls in reading at all ages, generally larger among the youngest students, in terms of both achievement and attitudes, and a gender gap in favour of boys in mathematics and physical science, again in terms both of achievement and attitudes but increasing with age. These results are supported by Blondin and Lafontaine (2005) using IEA TIMSS data (grades 4, 8 and 12) in mathematics and science.

This general pattern has however been nuanced. Researchers have found that the magnitude of the gender differences may depend on:

1. The test characteristics, including
    (a) The test content, such as the sub-domains assessed, the type of tasks or processes measured, the cognitive demands of the items, and the context of the questions. For instance in reading the largest gender differences have been reported for the narrative sub-domain (Elley 1992) or in continuous texts (Kirsch et al. 2002; Lafontaine and Monseur 2009a); in mathematics, researchers have reported males' better results in geometry (Hyde et al. 1990) and spatial representation (Johnson 1996; Mullis et al. 2000; OECD 2004), and females' in arithmetic and computation (Johnson 1996; Mullis et al. 2000). In science, the largest gender differences have been found in physical science (Comber and Keeves 1973; Johnson 1996; Mullis et al. 2000), and smaller ones in biology or chemistry.
    (b) The test format, such as the delivery of the test (paper-based versus digital) the format of the stimulus (continuous texts versus non-continuous texts), and the item format (multiple choice versus free response). A common finding concerning the item format is that "relative to males, females perform better on constructed-response than on multiple-choice items" (Bennet 1993: p. 20).
2. The study design, including the characteristics of the target population (age, educational level); the sample (representing the whole population or not), and the year of administration. From the meta-analysis on US data by Hyde and colleagues, it is clear that the nature of the sample influences the results. They showed that the selectivity (selected versus whole population) or the precocity (gifted students) of the student samples used to examine gender differences in mathematics had led to the persistent idea that boys consistently outperformed girls, whereas this difference in fact only appeared in high school, and girls were even slightly better at primary and lower secondary levels (Hyde et al. 1990).
3. The statistical analyses performed. This aspect is discussed in the next section.

Gender differences according to computed statistics

The growing success of using meta-analysis to report between-groups differences has non-negligible consequences on the kind of results reported. As pointed out by Feingold (1992), the success of meta-analysis at summarising research findings, accumulating effect sizes from different surveys or studies, leads to attention being focused on the central tendency statistics. Although Hedges and Friedman (1993) have shown that meta-analysis could be used to accumulate results in all facets of the scores' distribution, it has to be recognised that many authors have been more interested in the central tendency statistics and that the gender differences in variation and at the extreme tails of the distribution have not received the same amount of attention.

More importantly, effect sizes on gender differences have often being judged according to Cohen's (1977) guideline. As gender differences have generally been seen as small according to this classification (Nowell and Hedges 1998), this could have led to an overoptimistic evaluation of the actual gender gap. In this context, the work on U.S.-representative national samples by Hill et al. (2008)—arguing that policy-relevant gaps such as the gender differences in student achievement had to be estimated from empirical benchmarks—gives new insight to interpret gender differences data. Hill and colleagues showed in the U.S. context that gender differences do vary according to the outcome measured (larger differences in reading than in mathematics) and according to the educational level of the students. In reading, the gender gap increases with age, in favour of females: the effect size is −0.18 at grade 4, −0.28 grade at 8 and −0.44 at grade 12. In mathematics, the effect size in favour of males is 0.08 at grade 4, 0.04 at grade 8 and 0.09 at grade 12.

However, to address policy-relevant gender gaps (such as the issue of the overrepresentation of boys in the lowest proficiency levels in reading or the underrepresentation of girls in math and science in higher education), we need to pursue Hill et al.'s (2008) work, providing empirical benchmarks at different proficiency levels on representative samples.

Extreme tails

In the case of normal distribution and homogeneity of variance, differences at the upper and lower tails of the distribution may logically be inferred from the mean difference. However, in case of non-homogeneity of variance, gender differences vary by proficiency level; looking at extreme tails helps to nuance the outcomes on gender differences. For instance, Halpern and colleagues (2007, p. 40) conclude their review on sex differences in science and mathematics by arguing that "substantial evidence suggests that the male advantage in mathematics is largest at the upper end of the ability distribution".

For reading, Wagemaker (1996, p. 42) noted, based on IEA Reading Literacy Study data, that "in some countries, it is evident that disparity between boys and girls is not uniformly systematic across the ability distribution".

It is worth noting that, while boys' underachievement in reading is of growing concern, little education research has addressed this issue in terms of gender differences at the lower tail of the distribution, compared to those focusing on boys' better achievement at the higher end of the distribution in mathematics. According to Halpern et al. (2007), one of the reasons for this "higher end" focus could be the importance of quantitative abilities in many occupations. However, low abilities in reading are also a serious handicap to gaining access to the labour market (OECD and Canada 2000).

Gender variability

Although gender differences in variability are not a core concern of research into gender differences in achievement, studies in this area have a long tradition. According to Feingold (1992) the "greater male variability hypothesis" was formulated by Ellis as early as 1894. Based on psychological, medical and anthropological evidence, Ellis had noticed that males consistently presented greater variability than females.

In the field of education, studies including variability analyses by gender tend to confirm the greater male variability hypothesis (Beller and Gafni 1996). Others report more balanced results. In their review of sex differences in variability, Maccoby and Jacklin (1974) concluded that males were more variable than females in mathematical and spatial abilities and equally variable in verbal ability. Feingold (1992, p. 74) consistently found that “males were more variable than females in general knowledge, mechanical reasoning, quantitative ability, spatial visualisation, and spelling. There was essentially homogeneity of variance for most verbal tests, short term memory, abstract reasoning and perceptual speed”.

Purpose of the study

We explore gender differences in reading, mathematics and science since the 1990s (i.e. from the decade where co-education is achieved in most industrialized countries), using international large-scale survey data. We argue with Feingold (1992) that contemporary research on gender differences in intellectual abilities has focused on male–female differences in average performance, implicitly assuming homogeneity of variance. We also argue with Hill et al. (2008) that we need empirical benchmarks to assess policy-relevant gaps in students’ achievement, such as the gender gaps. We finally argue that these benchmarks have to be computed at different levels of the achievement distribution, since interventions to reduce the gender gap often focus on specific target groups: the least or the most proficient students.

We computed variance ratios and effect sizes for the mean and the percentiles 5, 10, 90 and 95. Our goal was to provide evidences from many national representative samples to answer the following questions:

1. To what extend do gender differences at the extreme tails of the distribution vary compared to gender mean differences, and in which direction?
2. Does the gender gap vary according to the content area and to the educational level of the students?
3. Is “greater male variability” still observed in recent studies of cognitive skills in reading, mathematics and science?

Methods

Data

The IEA and OECD PISA surveys were selected because they are the most ambitious in terms of world-wide coverage—with about five decades of experience in the case of the IEA—and provide quality databases and well-documented technical reports. International datasets are electronically available for six IEA studies: Progress in International Reading Literacy Study (PIRLS) 2001, 2006, 2011; Trends in International Mathematics and Science Study (TIMSS) 1995, 1999, 2003, 2007 (all the databases are available on the IEA Study Data Repository: http://rms.iea-dpc.org); and for five PISA cycles: PISA 2000, 2003, 2006, 2009, 2012 (available on the OECD website: http://www.pisa.oecd.org/). Methodological information is available in the technical reports on each survey (Adams and Wu 2002; Martin et al. 2000; Martin and Kelly 1996, 1997; Martin and Mullis 1996, 2012; Martin et al. 2004; Martin et al. 2003, 2007; OECD, 2005, 2009a, 2014b; Olson et al. 2008).

Analysis

Data from every country or education system in these international databases were included in this study, even those flagged in the international reports. For TIMSS, when two grade-based populations were assessed (generally referred to as Population I at the primary level, and Population II at the secondary level), both were included. However, the Population III that was assessed in TIMSS 95 and corresponded to the final year of secondary education was not included, because at this educational level school is no longer compulsory in many education systems, which makes the national samples more selective (moreover, gender differences in mathematics for population III in TIMSS 1995 have been analysed by Mullis and Stemler 2002). For PISA, the three content areas (i.e. reading, mathematics and science literacy) were used for each data collection cycle. In total, the analyses included 1654 cases, each of which corresponded to the assessment of one content area in one population in one country.

For any statistics reported in this study, the statistic and its corresponding sampling variance were computed five times, one on each plausible value, and aggregated according to the methodology presented in IEA/OECD technical reports and database manuals: the standard errors for IEA databases were estimated using the 75 JK2 replicates weights and the standard errors for PISA databases were estimated using the 80 Fay replicate weights.

Two types of statistics were computed:

1. Effect sizes for the means, and at percentiles 5, 10, 90 and 95. Although effect sizes have often been used on central tendency statistics only, Hedges and Friedman (1993) have shown that they could also be used to compare the magnitude of the gender difference at different portions of the score distribution. As the denominator for computing the effect size, we used the pooled standard deviation, as recommended in the *PISA Technical Report* (OECD 2009b, p. 195)

$$\frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\frac{\sigma_2^1 + \sigma_2^1}{2}}}$$

   Females were considered the target group. A positive value indicates a gap in favour of females, a negative value a gap in favour of males. The effect sizes were computed for each participating country for each survey, and then averaged across countries (each country being equal to one). We did not compute standard errors for the effect sizes. These were mainly computed for illustrating the gender gap at different levels of the proficiency distribution and being able to compare gender differences across studies. They would have been useful in order to test the significance of some of the results (trends for instance). However, one of the aims of this paper was to be exhaustive and to include all countries participating in international studies from the middle of the nineties. Testing differences in gender gaps according to the content area or the education level, for instance, would have required limiting the comparisons to the few countries that participated to all surveys included in this study which would have reduced in another way the scope of this paper.

The variance ratio. Differences in gender variability were measured by the variance ratio. This ratio has been widely used by authors examining gender variability (Feingold 1992; Nowell and Hedges 1998). The variance ratio expresses the variance of males in relation to that of females. A variance ratio greater than 1 means that males' variance is higher than females' variance. A mean variance ratio has been computed for each study (each country being equal to one). For each study, we computed also (1) the proportion of participating countries where the variance ratio is significantly greater than 1 (proportion of countries where the variance is significantly greater for males); (2) the proportion of countries where the variance ratio is greater than 1, but not significantly; (3) the proportion of countries where the variance ratio is lower than 1, but not significantly, and (4) the proportion of countries where the variance ratio is significantly lower than 1 (variance significantly greater for females).

Results

Effect sizes at the extreme tails of the score distribution

Figure 1 shows the gender gaps in reading, mathematics and science, by education levels. Detailed results (by survey) are presented in Table 1. In reading, the effect size on the mean at the primary level is 0.23, and is much larger at the secondary level (0.40), both effects being in favour of females. In mathematics and science, the effect sizes on the means are much lower. In mathematics, the effect size is −0.04 with international studies focusing on primary education, and is −0.07 for studies involving students in secondary education. In science, the male advantage is −0.05 for the youngest students and −0.07 for the oldest.

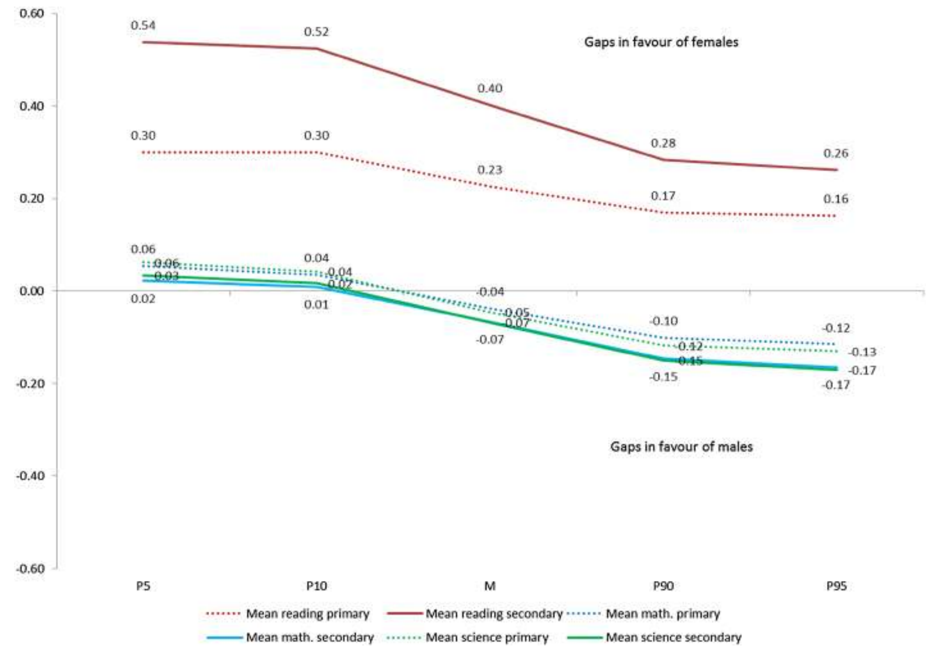

Fig. 1 Gender effect sizes in mean and extreme tails of the distribution in reading, math. and science. Gender effect sizes at percentiles 5, 10, 90, 95 and on the mean scores, by content area and education level. Positive values indicate higher scores for females; negative values indicate higher scores for males. *Data* IEA (TIMSS/PIRLS) and OECD (PISA) surveys, 1995–2012

Table 1 Gender effect sizes in mean and extreme tails of the distribution in in reading, math. and science, by survey

Content	Level	Orga.	Study	*N*	P5	P10	M	P90	P95
Reading	Primary	IEA	PIRLS 2001	*36*	0.34	0.33	0.25	0.20	0.20
			PIRLS 2006	*45*	0.28	0.27	0.21	0.16	0.15
			PIRLS 2011	*57*	0.28	0.30	0.22	0.15	0.14
	Secondary	OECD	PISA 2000	*42*	0.47	0.46	0.35	0.24	0.22
			PISA 2003	*41*	0.52	0.50	0.36	0.25	0.23
			PISA 2006	*56*	0.56	0.54	0.42	0.30	0.27
			PISA 2009	*65*	0.56	0.56	0.45	0.33	0.31
			PISA 2012	*68*	0.58	0.56	0.43	0.30	0.28
Math.	Primary	IEA	TIMSS 1995 P1 L	*24*	0.00	−0.01	−0.08	−0.14	−0.16
			TIMSS 1995 P1 U	*26*	0.06	0.04	−0.03	−0.08	−0.10
			TIMSS 2003 P1	*29*	0.07	0.05	−0.02	−0.09	−0.10
			TIMSS 2007 P1	*44*	0.09	0.06	−0.02	−0.09	−0.10
	Secondary	IEA	TIMSS 1995 P2 L	*39*	0.00	−0.01	−0.07	−0.14	−0.16
			TIMSS 1995 P2 U	*41*	0.01	−0.01	−0.08	−0.15	−0.15
			TIMSS 1999 P2	*38*	0.03	0.01	−0.05	−0.12	−0.13
			TIMSS 2003 P2	*51*	0.07	0.07	0.00	−0.07	−0.09
			TIMSS 2007 P2	*57*	0.13	0.12	0.03	−0.05	−0.07
		OECD	PISA 2000	*42*	−0.01	−0.02	−0.09	−0.17	−0.19
			PISA 2003	*41*	0.01	−0.01	−0.11	−0.21	−0.24
			PISA 2006	*57*	−0.01	−0.03	−0.10	−0.19	−0.21
			PISA 2009	*65*	−0.01	−0.02	−0.10	−0.18	−0.20
			PISA 2012	*68*	0.01	0.00	−0.09	−0.18	−0.21
Science	Primary	IEA	TIMSS 1995 P1 L	*24*	0.02	0.00	−0.09	−0.17	−0.18
			TIMSS 1995 P1 U	*26*	0.01	0.00	−0.10	−0.18	−0.20
			TIMSS 2003 P1	*29*	0.10	0.08	0.00	−0.06	−0.07
			TIMSS 2007 P1	*44*	0.12	0.09	0.01	−0.06	−0.07
	Secondary	IEA	TIMSS 1995 P2 L	*39*	−0.14	−0.16	−0.23	−0.30	−0.33
			TIMSS 1995 P2 U	*41*	−0.14	−0.17	−0.24	−0.32	−0.34
			TIMSS 1999 P2	*38*	−0.07	−0.10	−0.18	−0.25	−0.27
			TIMSS 2003 P2	*51*	−0.03	−0.04	−0.10	−0.16	−0.18
			TIMSS 2007 P2	*57*	0.14	0.14	0.04	−0.06	−0.07
		OECD	PISA 2000	*42*	0.12	0.11	0.01	−0.07	−0.09
			PISA 2003	*41*	0.07	0.05	−0.04	−0.13	−0.15
			PISA 2006	*57*	0.11	0.09	0.01	−0.08	−0.10
			PISA 2009	*65*	0.14	0.13	0.04	−0.05	−0.07
			PISA 2012	*68*	0.14	0.12	0.02	−0.08	−0.10
Mean reading				*410*	0.45	0.44	0.34	0.24	0.23
Mean math.				*622*	0.03	0.02	−0.06	−0.13	−0.15
Mean science				*622*	0.04	0.02	−0.06	−0.14	−0.16
Mean primary				*384*	0.12	0.11	0.03	−0.03	−0.04
Mean secondary				*1270*	0.13	0.12	0.03	−0.06	−0.08
Mean IEA				*836*	0.07	0.05	−0.03	−0.09	−0.11
Mean OECD				*818*	0.22	0.20	0.10	0.01	−0.02
Mean reading primary				*138*	0.30	0.30	0.23	0.17	0.16
Mean reading secondary				*272*	0.54	0.52	0.40	0.28	0.26
Mean math. primary				*123*	0.06	0.04	−0.04	−0.10	−0.12
Mean math. secondary				*499*	0.02	0.01	−0.07	−0.15	−0.17
Mean science primary				*123*	0.06	0.04	−0.05	−0.12	−0.13

Table 1 continued

Content	Level	Orga.	Study	*N*	P5	P10	M	P90	P95
Mean science secondary				*499*	0.03	0.02	−0.07	−0.15	−0.17
Mean				*1654*	0.13	0.11	0.03	−0.05	−0.07

A positive value indicates a higher score for females; a negative value indicates a higher score for males

N number of participating countries, *P5, P10, M, P90, P95* gender effect sizes at percentiles 5, 10, 90, 95 and for the mean scores

The effect sizes computed at the extreme tails of the distribution show that the size of the gender gap varies also according to the proficiency level of the students. In reading, even if the gender differences were fairly large along the entire proficiency distribution, they were particularly large at the lower tail, since effect sizes were sometimes about twice as large as at the upper tail. The largest gap in reading was observed on PISA 2012 (0.58) for the weakest students (percentile 5). The phenomenon was more pronounced in PISA and/or at the secondary level of education: the available database did not make it possible to distinguish the effect of educational level from the survey effect. At primary level, only IEA data were available, while at the secondary level, only PISA data were available.

In mathematics, effect sizes were smaller than in reading, but again, the size of the effect varies according to the proficiency level of the students. At the lower tail of the distribution, effect sizes were close to zero or in favour of females, while systematically at the upper tail, males were more proficient. The largest gap in mathematics was observed on PISA 2003 (−0.24) for the most proficient students (percentile 95). This tendency was more pronounced at the secondary level of education and in the PISA surveys. In the IEA Population II studies, the tendency for males to outperform females at the upper end of the distribution decreased across time.

Science results appear similar to mathematics results on Fig. 1, but looking at the data by survey (Table 1) reveals a situation somewhat more complex. At the primary level, a slight tendency for girls to be more proficient at the lower tail and for males to be more proficient at the upper tail was observed. At the secondary level, the gender difference in favour of males observed up to TIMSS 1999, at the lower and at upper tail of the distribution, has changed since the year 2000 in both the IEA and PISA surveys: at the lower tail, girls tend to perform somewhat better than males, and the male advantage at the upper tail tends to fade away across time.

Gender differences in variability

Table 2 focuses on gender variability. Four categories are presented: (1) the proportion of countries where the gender variance ratio was significantly greater than 1 (i.e. males' variance is significantly greater than females' variance); (2) the proportion of countries where the gender variance ratio is greater than 1, but not significantly; (3) the proportion of countries where the gender variance ratio is lower than 1, but not significantly; (4) the proportion of countries where the gender variance ratio is significantly lower than 1 (i.e. females' variance is significantly greater than males' variance). For each study, the mean of the country variance ratios and its standard error is also provided.

Table 2 Gender differences in variance ratios in reading, mathematics, and science

Content	Level	Orga.	Study	*N*	>1* (%)	>1 (%)	<1 (%)	<1* (%)	VR	(SE)
Reading	Primary	IEA	PIRLS 2001	*36*	25	58	17	0	1.09	(0.018)
			PIRLS 2006	*45*	22	69	9	0	1.08	(0.019)
			PIRLS 2011	*57*	25	70	5	0	1.10	(0.009)
	Secondary	OECD	PISA 2000	*42*	60	36	0	0	1.16	(0.013)
			PISA 2003	*41*	68	32	0	0	1.20	(0.012)
			PISA 2006	*56*	73	25	0	0	1.20	(0.010)
			PISA 2009	*65*	83	12	0	0	1.18	(0.008)
			PISA 2012	*68*	84	12	0	0	1.22	(0.008)
Math.	Primary	IEA	TIMSS 1995 P1 L	*24*	25	67	8	0	1.10	(0.020)
			TIMSS 1995 P1 U	*26*	23	69	8	0	1.10	(0.015)
			TIMSS 2003 P1	*29*	41	48	10	0	1.10	(0.013)
			TIMSS 2007 P1	*44*	39	55	7	0	1.12	(0.012)
	Secondary	IEA	TIMSS 1995 P2 L	*39*	26	67	8	0	1.10	(0.016)
			TIMSS 1995 P2 U	*41*	24	54	22	0	1.12	(0.018)
			TIMSS 1999 P2	*38*	29	63	5	3	1.12	(0.015)
			TIMSS 2003 P2	*51*	37	57	6	0	1.11	(0.011)
			TIMSS 2007 P2	*57*	47	40	12	0	1.14	(0.010)
		OECD	PISA 2000	*42*	31	62	0	0	1.12	(0.015)
			PISA 2003	*41*	56	37	7	0	1.17	(0.012)
			PISA 2006	*57*	54	37	7	0	1.13	(0.009)
			PISA 2009	*65*	55	38	3	0	1.12	(0.008)
			PISA 2012	*68*	63	35	0	0	1.15	(0.008)
Science	Primary	IEA	TIMSS 1995 P1 L	*24*	21	79	0	0	1.13	(0.019)
			TIMSS 1995 P1 U	*26*	35	62	4	0	1.15	(0.016)
			TIMSS 2003 P1	*29*	34	55	10	0	1.11	(0.007)
			TIMSS 2007 P1	*44*	39	48	14	0	1.12	(0.009)
	Secondary	IEA	TIMSS 1995 P2 L	*39*	33	59	8	0	1.12	(0.015)
			TIMSS 1995 P2 U	*41*	29	54	17	0	1.15	(0.020)
			TIMSS 1999 P2	*38*	32	61	8	0	1.14	(0.019)
			TIMSS 2003 P2	*51*	31	55	14	0	1.10	(0.012)
			TIMSS 2007 P2	*57*	49	37	14	0	1.15	(0.016)
		OECD	PISA 2000	*42*	36	62	0	0	1.15	(0.016)
			PISA 2003	*41*	56	41	0	0	1.15	(0.012)
			PISA 2006	*57*	67	28	7	0	1.14	(0.008)
			PISA 2009	*65*	71	25	3	0	1.15	(0.008)
			PISA 2012	*68*	75	21	0	0	1.17	(0.008)
Mean reading				*410*	58	37	5	0	1.15	
Mean math.				*622*	42	49	8	0	1.12	
Mean science				*622*	47	45	8	0	1.14	
Mean primary				*384*	30	61	9	0	1.11	
Mean secondary				*1270*	54	39	7	0	1.15	
Mean IEA				*836*	33	57	10	0	1.12	
Mean OECD				*818*	64	32	4	0	1.16	
Mean reading primary				*138*	24	67	9	0	1.09	
Mean reading secondary				*272*	75	21	3	0	1.19	
Mean math. primary				*123*	33	59	8	0	1.11	
Mean math. secondary				*499*	45	47	8	0	1.13	

Table 2 continued

Content	Level	Orga.	Study	*N*	>1* (%)	>1 (%)	<1 (%)	<1* (%)	VR	(SE)
Mean science primary				*123*	33	59	8	0	1.13	
Mean science secondary				*499*	51	41	8	0	1.14	
Mean				*1654*	48	44	7	0	1.14	

N number of participating countries, *>1** proportion of countries where the variance ratios are significantly greater than 1 (significantly greater for males), *>1* proportion of countries where the variance ratios are greater than 1, but not significantly, *<1* proportion of countries where the variance ratios are lower than 1, but not significantly, *<1** proportion of countries where the variance ratio significantly lower than 1 (significantly greater for females), *VR* mean variance ratio (*>1* greater for males), and its standard error (SE)

In 93 % of the 1654 cases, variance ratios are greater than one, which means that males' variance is larger than females'. Males' results are more widespread than females' results. The difference is statistically significant in 48 % of cases. This pattern is found whatever the content area (reading, mathematics and science), the educational level (primary/secondary), the year of the survey, or the study sample design (grade-/age-based samples). In only two of the 1654 cases is the variance of the female population significantly higher than the variance of the male population. The variance ratios are lower than one but not significantly in 7 % of cases, which means that on those few occasions, female variance is larger in absolute terms than male variance.

As can be seen in Table 2, the general pattern of greater variance for males changes, sometimes substantially, according to the domain or the educational level or between the IEA and the OECD. For instance, in reading, male variance is significantly greater than female variance in 238 (or 58 %) of the 410 cases. This proportion is larger than that observed for science (49 %) and for mathematics (42 %). Males at the secondary level more often (in 54 % of the cases) present a significantly larger variability than males at the primary level of education (30 % of significant gender difference at this level). In terms of the agencies organising the surveys, PISA surveys present much more variance ratios greater than 1 (64 %) than the IEA surveys (33 %).

The high proportion of variance ratio greater than 1 does not inform how much larger the variance for the male subpopulation is compared to the variance of the female subpopulation. On average, for all studies and countries, the variance ratio is 1.14. This means that on average, male variance is 14 % higher than female variance. Variance ratios range from 1.08, for PIRLS 2006, and 1.22, for PISA 2012, again in reading.

The variance ratios do not change much according to the content assessed, nor according to the agency organising the survey. There is almost no difference between educational levels in science and mathematics. In reading, however, the mean variance ratio at the primary level is 1.09 and increases to 1.19 at the secondary level.

Looking at the year of the surveys, no clear trend appears, either by content area or organisation.

One main finding emerges from this analysis: there are almost no exceptions to the higher male variance. The differences between content areas, educational levels, organisations, and surveys are quite slight, except for the difference between PISA and PIRLS in reading. One can just notice that the smallest variance ratio is found in primary reading, computed on PIRLS data, while the largest is also in secondary reading, computed on PISA data. This result might suggest that in reading the gender difference in variability increases with student age.

Nowell and Hedges (1998) found a strong correlation (0.74) between the variance ratio and the effect sizes of the mean gender difference. We also computed this correlation. With the data used in this study, the correlation is 0.42. It is worth noting that it ranges from 0.50 (for the correlation between the variance ratio and percentile 5) to 0.31 (for the correlation between the variance ratio and percentile 95). This indicates that the more males' scores vary compared to females' scores, the larger the difference between males and females at the lower end of the distribution.

Discussion

Consistent with Hill et al.'s (2008) results, effect sizes on the means differ for different types of outcomes and for different levels of education. At the international level, gender differences in reading are higher than gender differences in mathematics and science, which was also found for the U.S. with NAEP data (Hill et al., 2008). In reading, gender differences increase with age, which was also found by Hill et al. (2008). In mathematics and science, gender differences increase slightly with age, which was unclear in U.S. data. Yet, the values of effect sizes found with international data are pretty close to those found with U.S. data.

This study has also shown that reporting gender differences from central tendency statistics only is misleading, because gender differences at the extreme tails of the distribution can be quite different from what is observed with central tendency indices. In mathematics, the effect sizes of the mean gender difference are on average close to zero. However, a thorough analysis of the distributions reveals that males consistently outperform females at the highest levels. In science, the effect sizes of the mean gender difference are also close to zero, at both the primary and secondary level of education (with the exception of TIMSS 2003, population II). Nevertheless, at the higher end of the score distribution, an advantage for boys is consistently found. As in mathematics, male advantage in science is about twice as high at the upper end of the distribution than on the mean. In reading, while females' higher performance is already noticeable from the central tendency statistics, examining the lower tail of the distribution shows that the high percentage of low-performing boys is far more of an issue than the relative lack of highly proficient male readers. The effect size of the mean gender difference is equal to 0.34 on average, in favour of females. This gender difference is important enough to focus the attention on boys' underachievement in reading. However, looking at the tails of the distribution helps us to understand better what is going on. At the higher end of the score distribution in reading, there is still a gender difference in favour of females, but this is slightly reduced compared to the mean effect sizes (for instance, for percentile 95, the average effect size is 0.23). At the lower end of the distribution, the magnitude of the effect sizes is far more striking. At percentile 5, it averages at 0.45, ranging from 0.28 at the primary level to 0.56 at the secondary level. It is worth adding that Johnson (1996) and Elley (1992) found that the gender difference in reading decreased between primary and secondary education. The data we used are consistent with Hill et al.'s (2008) results, showing that nowadays the gender gap in reading increases with age. This means that male underachievement in reading is mostly due to the underperformance of the weakest males. It also means that most readers at risk are males and that their situation is much more problematic than that of females.

Concerning the variance, the results confirm what Feingold (1992) called the "unexpected stability of the male variability". It is worth highlighting that there are no exceptions to the greater male variance ratio: in nearly all cases, males present higher variability than females. This conclusion confirms the findings of Nowell and Hedges (1998) with regard to the U.S. 12th grade population, where a greater variance for the males was observed in almost all cases.

The greater male variance does not depend on the content area, the educational level and the survey characteristics, even if the greater male variance is slightly higher in reading than in mathematics and science at the secondary level rather than at the primary level, and in PISA surveys compared to IEA surveys.

As did Nowell and Hedges (1998), we also found a correlation between the variance ratio and the effect sizes of the mean gender difference, indicating that largest differences in variance are associated with larger mean differences.

Limitations

Some trends have been observed, as well as some evolution by educational levels. However, conversely in mathematics and in science, these results in reading are based on surveys conducted by two agencies with different definitions of the target population (age- versus grade-based sample). With PIRLS 2006 and PISA 2009 data, we compared gender variability and differences at the extreme tails for the whole samples and for subsamples of expected age and grade. From this analysis, we conclude that the population definition does not explain the large difference between IEA and PISA in reading. Analyses available on request show also that the kind of countries (industrialized countries versus developing countries) involved in the surveys did not influence the proportion of countries where males' variance ratios are significantly larger than females'. Therefore, more work is required to disentangle the effect of the methodological choices to fully understand and reach firm conclusions about the cross-time trends of gender differences in variability and at the extreme tails of the distributions.

This study only focuses on the description of the gender differences; it does not try to understand the origin or the consequences of the larger variability of the male distribution. It would be worth investigating non-cognitive outcomes such as engagement, intrinsic or instrumental motivation, and self-efficacy in the same perspective.

Conclusions

This study examined gender differences in variance and at the extreme tails of the score distribution in reading, mathematics and science. Ten databases from IEA and OECD PISA surveys were used to analyse such gender differences in an international perspective since 1995. The main results may be summarised in three points. (1) Gender differences vary by content area, students' educational levels, and students' proficiency levels. The gender differences at the extreme tails of the distribution are often more substantial than the gender differences on the mean. (2) Exploring the extreme tails of the distributions shows that the situation of the weakest males in reading is a significant issue. In mathematics and science, males are more often the best-performing students. (3) The "greater male variability hypothesis" is confirmed.

These findings are of key importance for ensuring gender equity. On the one hand, the situation of males at the lower end of the reading distribution is a matter of particular concern, since reading is a key and basic competence in educated societies. On the other hand, in mathematics and science, males perform better than females at the upper end of the distribution. Although the magnitude of this difference could be described as "small" (according to Cohen's categorization), it should be considered carefully, in view of the lower proportion of women in science and mathematics in college and in corresponding professional careers.

These findings of differentiated results at the extreme tails of the distribution in different academic content areas call for overall consideration of the way that males and females are grouped, tracked, retained, oriented and selected in education systems. Such an inquiry would have to include individual teachers themselves, since some researchers have found that scripts were assessed differently according to the gender of the student. For instance, Lafontaine and Monseur (2009b) have shown with an experimental study and with PISA data that high achievers in mathematics tend to be overestimated when males, and underestimated when females, and, conversely, low-achieving males tend to be underestimated while low-achieving females tend to be overestimated. It is interesting to note that data on teachers' evaluations in mathematics support findings from external assessments in accordance with greater male variability and with the lowest proportion of females at the highest level of the math proficiency distribution. Further research is needed to extend these results to other content areas, particularly to reading, where teachers' gender-related expectations may differ substantially from their expectations in mathematics, and in physical sciences where teachers' gender expectations should go in the same direction as in mathematics.

The asymmetry of male and female achievement and the multiplicity of sources of gender inequality undoubtedly imply mixed and complex solutions. Achieving gender equity does not mean achieving some utopian notion of strict gender equality in all domains and situations, but deciding on which gender inequalities are unfair, and prioritising the detection and suppression of such unfair situations (EGREES 2005). The literature and our results suggest an improvement for the weakest and the 'average' girls over the last decades. Why then is gender equality still not achieved at the highest levels? This questions the role of education, and in particular the role of the school in maintaining gender inequalities, and in the orientation process. Are students still partly oriented on a gender basis in the most demanding courses? It is likely that an unknown proportion of girls may make the decision to not engage in science- and mathematics- related fields of study for prioritizing their future family. However, it is also likely that others do not consider these fields as options for girls. How many will be advised by their teachers to simply consider it as a possibility? Regarding the possibilities for boys, although many efforts are made to enhance the reading performances of the struggling readers, policy initiatives specifically designed to close the gender gap in reading are rare, or may not exist in some countries (Brozo et al. 2014). The reading gender gap along the achievement continuum, as well as the unacceptable situation of the weakest boys, is an argument for the institutionalization of "boy- friendly" curriculums. Brozo and colleagues

(2014) suggest working on the diversity of the reading material provided at school to retain boys' interest, supporting boys' use of digital texts and alternative media, involving significant male models in reading activities, and focusing on practices that promote boys' reading engagement.

Authors' contributions

The work represents extensive collaboration and discussion between AB and CM. Both authors participated in the development of the rationale for the study. AB wrote the manuscript and CM carried out the analysis. Both authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Received: 19 October 2014 Accepted: 10 December 2015
Published online: 19 January 2016

References

Adams, R., & Wu, M. (Eds.). (2002). *PISA 2000 technical report*. Paris: OECD.

Barro, J., & Lee, J.-W. (2001). International data on educational attainment: updates and implications. *Oxford Economic Papers, 3*, 541–563.

Baudelot, C., & Estabelet, R. (2007). *Quoi de neuf chez les filles? Entre stéréotypes et libertés*. Paris: Nathan.

Beller, M., & Gafni, N. (1996). The 1991 International Assessment of Educational Progress in Mathematics and Sciences: the gender differences perspective. *Journal of Educational Psychology, 88*(2), 365–377.

Bennet, R.E. (1993). On the Meanings of Constructed Response. In R. E. Bennet & W. C. Ward (Ed.), Construction versus choice in cognitive measurement. issues in constructed response, performance testing, and portfolio assessment (pp 1–27). Hove: Lawrence Erlbaum Associates.

Blondin, C., & Lafontaine, D. (2005) Les profils des filles et des garçons en sciences et en mathématiques. Un éclairage basé sur les enquêtes internationales. In: M. Demeuse & M. H. Straeten, J. Nicaise, A. Matoul (Eds) Vers une école juste et efficace (pp. 317–34). Brussels: De Boeck.

Brozo, W., Sulkunen, S., Shield, G., Garbe, Ch., Pandian, A., & Valtin, R. (2014). Reading, gender, and engagement. *Journal of Adolescent and Adult Literacy, 57*(7), 584–593.

Buchmann, C., & DiPrete, T. (2006). The growing female advantage in college completion: the role of family background and academic achievement. *American Sociological Review, 71*(4), 515–541.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

Comber, L.C. & Keeves, J.P. (1973). Science education in nineteen counties. Stockholm and New York: Almqvist & Wiksell and Wiley-Halsted Press.

Dubet, F. (2010). L'école "embarrassée" par la mixité. *Revue française de pédagogie*, *171*, 77–86.

Egrees. (2005). Equity in european educational systems: a set of indicators. *European Educational Research Journal, 4*(2), 1–151.

Elley, W. B. (1992). *How in the world do students read? IEA Study of Reading Literacy*. The Hague: IEA.

Feingold, A. (1992). Sex differences in variability in intellectual abilities: a new look at an old controversy. *Rev Educ Res, 62*(1), 61–84.

Halpern, D. F., Benbow, C. P., Geary, D. C., Gur, R. C., Hyde, J. S., & Gernsbacher, M. A. (2007). The Science of sex differences in science and mathematics. *Psychological Science in the Public Interest, 8*(1), 1–51.

Hedges, L. V., & Friedman, L. (1993). Gender differences in variability in intellectual abilities: a reanalysis of feingold's results. *Review of Educational Research, 63*(1), 94–105.

Hill, C., Bloom, H., Black, A., & Lipsey, M. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives, 2*(3), 172–177.

Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance—a meta-analysis. *Psychological Bulletin, 107*(2), 139–155.

Jacobs, J. (1996). Gender inequality and higher education. *Annual Review of Sociology, 22*, 153–185.

Johnson, S. (1996). The contribution of large-scale assessment programmes to research on gender differences. *Educational Research and Evaluation, 2*(1), 25–49.

Kirsch, I., de Jong, J., Lafontaine, D., McQueen, J., Mendelovits, J., & Monseur, C. (2002). Reading for change—performance and engagement across countries. Results from PISA 2000. Paris.

Lafontaine, D., & Monseur, C. (2009a). Gender gap in comparative studies of reading comprehension: to what extent do the test characteristics make a difference? *European Educational Research Journal, 8*(1), 69–79.

Lafontaine, D., & Monseur, C. (2009b). Les évaluations des performances en mathématiques sont-elles influencées par le sexe de l'élève ? *Mesure et évaluation en éducation, 32*(2), 71–98.

Lietz, P. (2006a). Issues in the change in gender differences in reading achievement in cross-national research studies since 1992: a meta-analytic view. *International Education Journal, 7*(2), 127–149.

Lietz, P. (2006b). A meta-analysis of gender differences in reading achievement at the secondary school level. *Studies in Educational Evaluation, 32*(4), 317–344.

Maccoby, E. E., & Jacklin, C. N. (1974). *The psychology of sex differences*. Stanford: Stanford University Press.

Marks, G. N. (2008). Accounting for the gender gaps in student performance in reading and mathematics: evidence from 31 countries. *Oxford Review of Education, 34*(1), 89–109.

Marry, C. (2003). Les paradoxes de la mixité filles-garçons à l'école. Perspectives internationales. Rapport pour le PIREF et conférence du 16 octobre 2003 au ministère de l'Éducation nationale. 2003. http://back.ac-rennes.fr/orient/egalchanc/rapmixite22103.pdf

Martin, M. O., Gregory, K. D., & Stemler, S. E. (Eds.). (2000). *TIMSS 1999 Technical Report: IEA's Repeat of the Third International Mathematics and Science Study at the Eighth Grade*. Chestnut Hill: Boston College.

Martin, M. O., & Kelly, D. L. (Eds.). (1996). TIMSS Technical Report: Volume I Design and Development. Chestnut Hill, MA: Boston College

Martin, M. O., & Kelly, D. L. (Eds.). (1997). TIMSS Technical Report. Volume II. implementation and analysis, primary and middle school years. Chestnut Hill, MA: Boston College

Martin, M. O., & Mullis, I. V. S. (Eds.). (1996). *TIMSS: quality assurance in data collection*. Chestnut Hill: Boston College.

Martin, M.O. & Mullis, I.V.S. (Eds.). (2012). Methods and procedures in TIMSS and PIRLS 2011. Chestnut Hill: TIMSS & PIRLS International Study Center, Boston College.

Martin, M. O., Mullis, I. V. S., & Chrostowski, S. J. (Eds.). (2004). *TIMSS 2003 technical report*. Chestnut Hill: Boston College.

Martin, M. O., Mullis, I. V. S., & Kennedy, A. M. (Eds.). (2003). *PIRLS 2001 technical report*. Chestnut Hill: Boston College.

Martin, M. O., Mullis, I. V. S., & Kennedy, A. M. (Eds.). (2007). *PIRLS 2006 technical report*. Chestnut Hill: Boston College.

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 (TIMSS)*. Chestnut Hill: Boston College.

Mullis, I. V. S., & Stemler, S. E. (2002). Analyzing gender differences for high-achieving students on TIMSS. In D. F. Robitaille & A. E. Beaton (Eds.), *Secondary analysis of the TIMSS data* (pp. 277–290). Dordrecht: Kluwer Academic Publishers.

Nowell, A., & Hedges, L. V. (1998). Trends in gender differences in academic achievement from 1960 to 1994: an analysis of differences in mean, variance, and extreme scores. *Sex Roles, 39*(1), 21–43.

OECD (Ed.). (2004). Learning for tomorrow's world—first results from PISA 2003. Paris

OECD. (2005). *PISA 2003 technical report*. Paris: OECD.

OECD. (2009a). *PISA 2006 technical report*. Paris: OECD.

OECD. (2009b). *PISA data analysis manual SAS®* (2nd ed.). Paris: OECD.

OECD. (2014a). *Education at a Glance 2014: OECD indicators*. Paris: OECD.

OECD. (2014b). *PISA 2012 technical report*. Paris: OECD.

OECD & Statistics Canada (2000). Literacy in the Information Age. Final Report of the International Adult Literacy Survey. Paris: OECD; Canada: Minister of Industry.

Olson, J. F., Martin, M. O., & Mullis, I. V. S. (Eds.). (2008). *TIMSS 2007 technical report*. Chestnut Hill: Boston College.

UNESCO. (2011). *The hidden crisis: armed conflict and education*. Paris: United Nations Educational, Scientific and Cultural Organization.

Wagemaker, H. (Ed.). (1996). *Are girls better readers?: Gender differences in reading literacy in 32 countries*. Amsterdam: International Association for the Evaluation of Educational Achievement.

Willingham, W. W., & Cole, N. S. (1997). *Gender and fair assessment*. Mahwah: Lawrence Erlbaum Associates.