Large-scale Assessments
in Education

Trends in gender gaps: using 20 years of evidence from TIMSS

Sabine Meinck and Falk Brese[*]

*Correspondence:
falk.brese@iea-hamburg.de
IEA, Überseering 27,
22297 Hamburg, Germany

Abstract

Potential differences in achievement between female and male students have always been an interesting topic in educational research, as well as having policy and economic implications. This study provides an overview of the so-called "gender gap" in mathematics and science knowledge, based on an in-depth analysis of both extremes of student ability distributions. Evidence underpinning debate on gender inequality in education can be explored by analyzing trends in these distributions over the last 20 years. This new approach to gender gap analysis shows that while the gender gaps that existed 20 years ago have persisted, gender equality in education has increased. The persistent trend of an overrepresentation of male students in the group of high-achievers in both mathematics and science is striking, but male and female students are often also unequally represented at the lower end of the ability distributions. Patterns differ between countries and cycles. In many countries, male students constitute the majority of the lower end of the ability distribution, while in others, more female students are failing to achieve, especially at grade eight. Some countries have shown a reversed inequality trend over the last two decades. With the proposed approach in analyzing gender gaps, differences at the tails of the achievement distributions can be investigated even if the gender distribution is skewed. Policymakers could make use of the approach to closely monitor the development of achievement gaps in their countries and initiate measures to tackle potential causes of inequity, leading to gender inequalities regarding educational achievement.

Introduction

Differences in achievement between female and male students, often termed the "gender gap", have always been of interest, not only in educational research, but also from a political and economic context (UNESCO 2015a; Hausmann et al. 2009). These differences are frequently seen as a matter of inequality (Klasen 2002). Achieving strict gender equality in all situations or domains may seem to be a utopian goal. However, laying the foundations of gender equity has become a political issue and is seen as a general measure of justice and fairness, especially in the education context (EGREES 2005).

At an international level, gender equality is of high importance, leading UNESCO to declare gender equality as one of the most important goals for education (UNESCO 2015b), and ultimately to incorporate this aim within the framework of sustainable development goals (United Nations 2018). International comparative research is addressing the issue of gender differences continuously, and the topic was prominent in

many recently conducted international large-scale assessments in education, including for example the 2015 TIMSS and PISA cycles (Mullis et al. 2016a; Mullis et al. 2016b; OECD 2016).

Literature review and theoretical framework

Gender equality and equity in education and society

Gender equality and equity in education is an issue under discussion for more than a century. At the time the right and obligation of schooling was introduced, single-sex schools dominated the educational landscape in many countries. Subjects taught to male and female students differed, reflecting the expected course of life of these children. Consequently, various subjects aimed at a certain gender group; for example cooking would be aimed at girls (Trueman 2015). Nowadays, fairly equal opportunities to learn have been established in the vast majority of countries for female and male students. However, the traditional patterns keep influencing in very powerful ways the life course of male and female students. For example, girls—as opposed to boys—still opt more for professions within the social sector and less often for sectors related to the so-called STEM (science, technology, engineering and mathematics) subjects. These patterns can be observed with career and study choices already prior to entering the work force (UNESCO 2017).

Moreover, stereotypes related to these traditional understandings of role models persist into the present, and they do influence what happens in the classroom today. Stereotypes affect professional action of teachers, parental influence and expectations, and consequently students' self-concept, decisions, actions, and achievement. For example, Nguyen and Ryan (2008) review existing literature regarding negative effects of stereotyping for girls in mathematics, and prove in an experiment the effect of stereotype threats on achievement. Retelsdorf et al. (2015) found negative associations between teachers' gender stereotype and boys' reading self-concept, disadvantaging male students in their reading achievement. They point out that stereotypes can explain the long-term development of self-concepts as a relatively stable personal characteristic. This, in relation with theories on self-fulfilling prophecy, may be one explanation of manifested differences over the course of schooling.

One aspect of gender differences receiving high attention is related to STEM education. UNESCO (2017) reports on girls' and women's education in STEM find that, to date, girls are still underrepresented in choosing STEM disciplines for studying and as their career paths. Also international comparative studies observe similar patterns. The IEA TIMSS-Advanced study on upper secondary students studying advanced mathematics and science conducted in 2015 found (far) more male students in these advanced courses in most of the participating countries (Mullis et al. 2016c). Further, male students on average achieve significantly higher than girls in—again—most of the countries.

However, countries have tried (in some cases since decades) to counteract gender inequity. UNESCO (2017) has compiled examples of the various kinds of interventions and programs regarding the gender differences in STEM education and outcomes. The list of examples comprises action targeted at the individual (female) students, for example single-sex workshops for girls to act as scientists led by same-sex tutors in the UK to facilitate girls' interest in STEM subjects and careers. Other examples are one-week

STEM camps (Kenya), where female students carry out experiments and visit companies offering STEM jobs, or "Science, Technology and Mathematics Education (STME) Clinics" (Ghana) which bring together girls in secondary schools with female scientists who could act as role models. Also on education-system level there was action taken: Improvements regarding school safety, education of teachers, smaller mathematics and science classes and better curriculum coverage could be identified (Mullis et al. 2016). Finally, at the level of the country or society itself, policies like quotas or financial investments to promote the image of women in STEM jobs are only some examples that have been implemented so far.

Measuring gender equality and equity in education

While the aforementioned studies aim towards showing causes of gender inequity in education, large-scale assessment data cannot provide this type of insight. Rather, figures of equality (i.e., comparing for example the achievement of gender groups) can be used as indicators of gender equity within educational systems, assuming there are no other factors determining differences between the sexes, such as genetic disposition. In other words, gender equity is understood in this context as a synonym for fairness and equal opportunities for female and male students, while gender equality represents an empirically measurable outcome of equity.

But what constitutes equality? As Allison (1978) states, already the choice of the measure to represent equality can make a difference regarding the perception of equality. Especially in cross-country comparisons, the question of what or who to compare to becomes a deciding factor on the resulting perception of the magnitude of inequality or even the lack of equality. In general terms, there are two types of comparisons.

First, comparisons could be based on absolute figures or distributions, setting or according to some standard across all compared entities, for example education systems. These standards could be related to various aspects of equality, like conditions, or outcomes. UNESCO, for example, classifies concepts for measuring equity in education accordingly (cf. 2018, 23ff.): 'Equality of conditions' would mean that conditions of education are the same for everyone. 'Equality of outcomes' entails minimum educational outcomes (e.g., a certain completed level of schooling) for everyone. When 'education is independent from personal characteristics', equality is related to the impartial implementation of education. A 'positive relation between education and ability' would be a kind of equality where students with higher ability are provided with higher quality education. This would be the case in countries with a tracked school system. Finally, a 'positive relation between education and being disadvantaged' (e.g., regarding some criteria like income) would provide higher quality, more or special focused education to disadvantaged students. In international large-scale studies in education, outcome measures often translate to certain benchmarks of achievement or proficiency levels (see, for example, Mullis et al. 2016, 61ff; Schulz et al., 2017, 44ff; OECD 2016, 59ff.). Gender equality is reported in relation to some standard or threshold applied to each country. While giving highly valuable information, this type of comparison has one disadvantage, that is, the overall achievement distributions vary greatly between countries, and there may be very few overlap at the ends of the distributions, leaving few room to compare gender equalities at these ends. For example, if one wishes to compare gender

ratios of eights grade students reaching the advanced TIMSS mathematics benchmark in 2015,[1] results are scanty, as more than two-thirds of the countries have less than 10% of their students reaching this benchmark; a significant number of countries even have no students at all in this category. Switching to the next lower benchmark level (the "high benchmark") will not help, as the majority of students in high-ranking countries achieve this category, hence, instead of focusing on the distribution tail, one would again rather focus on major population parts.

A different approach would be to compare achievement distributions across countries, using measures of relative equality within countries as comparison criterion. There are several implications of the latter approach: First, there is no need for (agreeing upon) a criterion acceptable and applicable across different education systems and countries. Second, the notion of comparability implies taking into account the country contexts. Looking at relative distributions in specific contexts or situations can provide more detailed insights into variation, especially in cases where international standards do not fit well. Subsequently, such analysis has the potential to reveal tailored options for addressing issues. For example, if gender gaps exist especially at the lower end of the achievement distribution, measures addressing weaknesses of low achieving students (tailored to the affected gender) will be more effective to tackle the issues.

The vast majority of related previous research based on international large-scale assessment data has focused on comparisons of the mean achievement of female and male students, only relatively few studies have addressed gender differences at different levels of achievement. Moreover, according to observations by Halpern et al. (2007), research focusing on gender differences at the tails of the ability distributions concentrates most on the upper tail (for example, Benbow and Stanley 1983). Only little research is concerned with the lower tail of the ability distribution. Research on the upper tail of the distribution looked at the absolute distribution of female and male students (for example, Hedges and Nowell 1995; Strand et al. 2006).

Baye and Monseur (2016) looked at differences in the variability of students' achievement using TIMSS, PIRLS and PISA data spanning more than 20 years. They point out that gender differences at the extreme tails of the achievement distribution are often more substantial than average differences. Males were more frequently among the highest performing students in mathematics and science, but male students also varied more than female students in their level of performance. Bergold et al. (2016) also identified a higher variability in the achievement of male students, with male students overrepresented as a group among both highest and lowest performing students. The study looked at 17 countries participating in TIMSS and PIRLS in 2011 with fourth-grade students, and at various achievement domains (reading, mathematics and science) simultaneously. Significant variations between country profiles have been acknowledged, suggesting that a single generic model of explaining gender differences may not be reasonable. Both publications (Baye and Monseur 2016; Bergold et al. 2016) include a comprehensive review of the literature related to gender differences in general and theories on the greater variability of males regarding achievement.

---

[1] http://timssandpirls.bc.edu/timss2015/international-results/timss-2015/mathematics/performance-at-international-benchmarks/performance-at-the-international-benchmarks-of-mathematics-achievement/.

Recently, some research used quantile regression analysis (Davino et al. 2014) as another way of differentiating inequalities along ability distributions. For example, Costanzo and Desimoni (2017) found varying gender inequalities for the different quantiles of the mathematics and reading scores distributions, using data from an Italian study of second and fifths grade students.

Objectives and research questions

The research presented in this paper was stimulated by the following considerations: (1) we acknowledge the fact that gender equity remains an important issue on the political agenda of many countries; (2) many countries have introduced measures tackling gender inequity over the past decades; (3) previous findings suggest that average achievement is not the most comprehensive indicator of gender equality, rather, unequal gender ratios can be observed especially at the tails of the achievement distributions. Consequently, we would like to expand current knowledge by adding a perspective on trends over time regarding gender differences at the tails of the ability distributions. Deviating from approaches used in previous research, we will implement a statistical analysis method that accounts for potentially skewed overall gender distributions within education systems. Focusing on relative ability distributions of female and male students rather than on absolute distributions, for example according to internationally defined proficiency levels, we will be able to identify gender inequalities better for countries where students' results do not show (enough) variation across these standardized levels. Further, we can compare countries with regard to gender differences even if the average achievement of students varies a lot between these countries. By analyzing the tails of the relative distribution, we will get information on gender inequalities for the highest and lowest achieving students. We argue that these more fine-grained results—compared to overall achievement averages—could be used to develop measures and policies tailored more specifically to these groups of students. Regarding equal opportunities as an aspect of equity, the lower tail of the achievement distribution might be of special interest, as extremely low performance can seriously affect future options in school and later on in life. Finally, the proposed method is robust to skewed overall distributions of female and male students. This is an important aspect when including countries into the analyses were school enrolment is gender-dependent, or for trend analysis if overall gender ratios change over time.

Using TIMSS data, we sought to evaluate whether differences between girls and boys regarding their mathematics and science achievement at fourth and eighth grade changed over the last 20 years. We considered four central research questions.

(1) Considering mathematics and science achievement of fourth and eighth graders, is there an equal gender distribution at the top and bottom end of the achievement distribution within participating countries?

(2) If there is a gender gap, did it change over time? More specifically, did gender gaps change at the fourth grade from 1995 to 2015, and similarly at the eighth grade?

(3) Looking at specific student cohorts in the fourth grade and again 4 years later at the eighth grade by following up the cohort, did the achievement gap widen or narrow?

Fig. 1 Identifying the tails of the achievement distributions
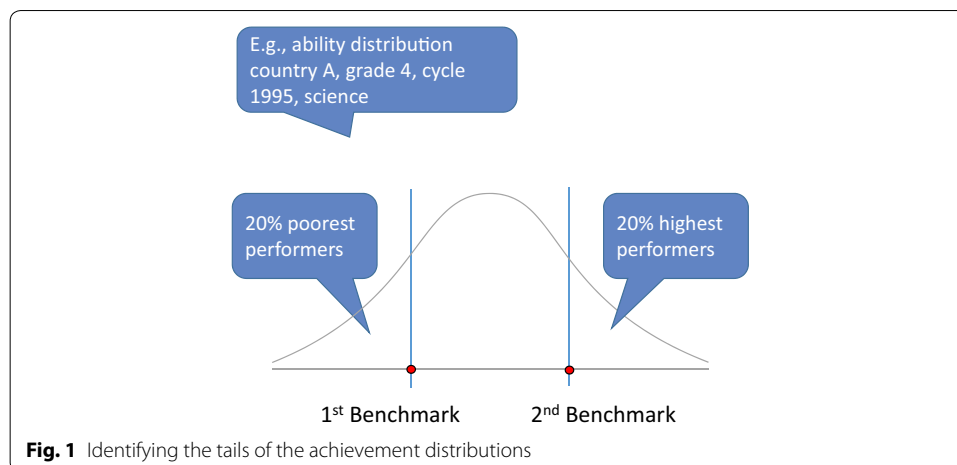
(4) Are these developments internationally generalizable or can different patterns be observed in different countries or groups of countries?

Data and methods

We analyzed data collected from education systems participating in TIMSS 1995, TIMSS 2015, and at least one additional intermediate cycle of TIMSS. Eighteen education systems at Grade 4 and 20 education systems at Grade 8 satisfied the requirements.

To enable a longitudinal analysis of specific cohorts at country level for addressing research question (3), we reduced the scope of our research further. From the countries included above, we chose only those who participated at grade four in 1995 and 2011, and at grade eight in 1999 and 2015. Consequently, we followed up two cohorts in 11 countries: students who attended grade four in 1995 and grade eight in 1999, and a cohort born 16 years later, with students attending grade four in 2011 and grade eight in 2015. It should be noted that only representative samples of the same cohorts were tested, and not the same students at different ages.

We first identified the 20% best and poorest performers in each country and cycle per grade and subject domain, using the overall mathematics and science achievement scores, by performing a percentile analysis. This analysis resulted in two benchmark scores per population, subject domain, country, and cycle, dividing the best and the poorest performing 20% from the remaining populations (see Fig. 1). Readers should be reminded that the achievement levels of these groups differ greatly among countries as shown in Fig. 2, but these differences are not of interest for this paper. Instead, our focus is purely on the gender gap within and across countries, and time.

In a second step, we estimated the differences in percentages of male and female students reaching or failing respective benchmarks resulting from the first step as illustrated in Fig. 3. Again, separate analyses were run for different grades, subject domains, countries and cycles. The results of these analyses were the relative distributions of female and male students in the groups of "high" and "low" performers. The percentages were computed in a way that allows a direct comparison between the relative

Fig. 2 Ability distributions vary greatly among countries

Fig. 3 Estimating relative distributions of female and male students in the groups of 'high' and 'low' performers
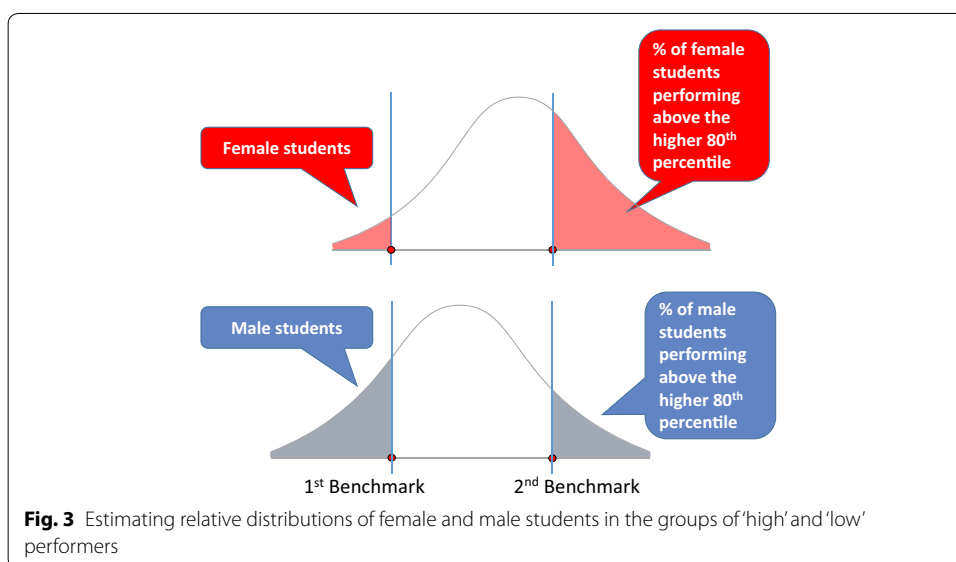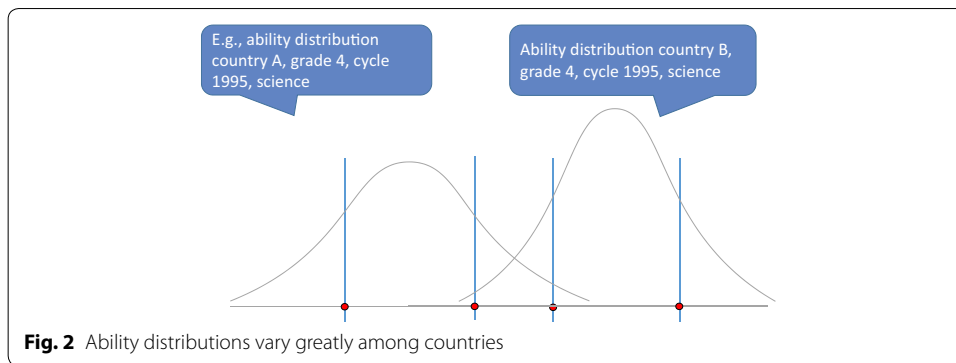
distributions even in populations with overall skewed gender distributions (i.e., populations with more female than male students comprising a grade, or vice versa).

We accounted for the complex sample and assessment design by using sampling weights for the estimation of population parameters, and applying jackknife repeated replication and plausible values for the estimation of standard errors (Foy 2017).

Results

General trends in gender gaps

Overall, there are more differences in trends for gender gaps between grades for the same subject than between subjects within grades. Gender gaps for mathematics achievement at grade four and their trends over time are more similar to those for science achievement at the same grade than trends at grade eight. In other words, it seems that, regarding gender differences, grade matters more than subject.

Trends in gender gaps in mathematics achievement at grade four

Twenty percent highest performing students (above 80th percentile)

Overall, there were relatively more boys than girls among the top 20% of students by achievement (Table 1). This applies to almost all countries and cycles. In about two-thirds of all observed cases, this difference was significant.

Kuwait was the only exception to this general finding; in 2007 and 2011, female students were significantly overrepresented in this group. Singapore was the only country to possess remarkable gender balance in all cycles from 1995 to 2015. Finally, only in Japan has inequality reduced, with an initially significant gender gap in favor of boys reducing from 2003 (and becoming insignificant in later cycles).

In 2015, in 12 out of 18 countries there was a significant gender gap favoring boys; the same tendency was observed in the other countries, but the differences were not significant. In seven countries (Australia, Czech Republic, Hong Kong (SAR), Hungary, New Zealand, Slovenia, and the United States), a gender gap in favor of boys widened over the last 20 years, starting from a small and mostly insignificant difference in 1995 and 2003 to a significant gap in 2015, posing questions surrounding potential causes of this apparent increase in inequality.

Twenty percent lowest performing students (below 20th percentile)

Overall, fewer significant gender gaps were observed in this group of low-performers (Table 1). Moreover, there were no generalizable trend patterns. In The Netherlands, female students were significantly overrepresented in this group compared to their male peers, a gap that has remained fairly constant over the last two decades. In Iran, Kuwait and Singapore, there were significantly more boys represented among these low-performing students in more recent cycles.

A gender gap existing in New Zealand in 1995 (again, with more male students being part of this group) became insignificant in all later cycles.

Trends in gender gaps in mathematics achievement at grade eight

Twenty percent highest performing students (above 80th percentile)

As with grade four, there are relatively higher percentages of boys than girls among the top 20% highest achieving eighth grade students in mathematics (see Table 2). Gaps predominantly favor boys, in all significant gaps but one (Thailand). In 1995, there was a significant gender gap in nine countries, whereas, in 2015, this was true for only five countries, showing a reduction somewhat of the gap among the countries considered.

Over the last two decades, relatively constant gender gaps favoring boys can be observed for Italy, Japan, Korea and the United States. In three countries (England, Iran and Israel), gender inequality decreased over the same period. Only in Thailand, there was a tendency to have relatively more girls in this group (gap significant only in 2007).

Twenty percent lowest performing students (below 20th percentile)

Overall, the gender gaps in mathematics achievement of the 20% lowest performing students in grade eight differ by country (see Table 2). Gender gaps are often smaller

Table 1 Gender gaps in percent of group membership related to mathematics achievement, grade 4

Percentile	Cycle	Australia	Cyprus	Czech Republic	England	Hong Kong (SAR)	Hungary	Islamic Republic of Iran	Ireland	Japan
Upper 20th	1995	-2.0	-3.9 *	-1.0	-4.0 *	-2.0	-2.1	-6.2 *	1.1	-3.9 *
	2003	-3.2	-5.6 *		-2.5	-2.0	-1.7	1.1		-5.1 *
	2007	-4.7 *		-5.6 *	-1.2	-3.9 *	-1.6	2.2	-2.6	-1.7
	2011	-3.9 *		-6.1 *	-4.1	-6.4 *	-3.7 *	-1.9	-2.7	-2.6
	2015	-5.2 *	-5.1 *	-3.4 *	-4.4 *	-6.6 *	-3.5 *	-0.3		-1.7
Lower 20th	1995	0.6	0.3	3.5 *	0.6	1.3	0.0	4.5 *	-4.2 *	-0.5
	2003	0.1	1.2		-1.7	-2.0	-0.6	-4.1		-1.3
	2007	-0.4		-0.2	-1.0	0.5	1.1	-7.0 *		-0.6
	2011	0.3		3.2	-1.6	1.2	-0.7	-2.0	0.5	0.4
	2015	1.4	0.3	2.0	0.0	2.1	0.9	-5.7 *	-0.1	-2.0

Percentile	Cycle	Republic of Korea	Kuwait	The Netherlands	New Zealand	Norway	Portugal	Singapore	Slovenia	United States
Upper 20th	1995	-5.0 *	-0.4	-4.2 *	-0.6	-4.1 *	-2.7 *	-0.2	-1.2	0.0
	2003			-2.3	-1.3	-3.5 *		0.4	-4.2	-4.5 *
	2007		6.9 *	-5.1 *	-3.5 *	-3.9 *		-0.7	-4.0 *	-3.5 *
	2011	-4.6 *	6.9 *	-3.7	-2.0	-6.0 *	-4.2 *	0.5	-6.0 *	-5.1 *
	2015	-4.9 *	-0.3	-4.8 *	-2.8 *	-0.4	-5.9 *	-0.2	-5.2 *	-3.3 *
Lower 20th	1995	4.8 *	-5.2	3.1 *	-5.7 *	3.7 *	1.5	-3.2 *	1.0	-1.2
	2003			3.4	-1.2	-0.6		-4.8 *	-1.1	1.6
	2007		-12.4 *	4.6 *	-2.2	1.8		-3.5 *	0.6	1.2
	2011	0.7	-11.9 *	4.2 *	-1.6	1.0	1.0	-2.6 *	2.1	2.0 *
	2015	1.8	-7.0 *	3.7 *	-1.8	-2.7	3.0	-2.0	-0.9	1.6

Positive gap values and red bars indicate higher relative percentages of girls in the group. In order to enable readers to correctly interpret Tables 1, 2, 3, 4, please see the following interpretation example for Table 1: "In Australia, the proportion of male students in the group of the twenty percent highest achievers in all cycles is higher than the proportion of female students in that group. This disproportional group allocation is significant in 2007, 2011 and 2015. In the group of the twenty percent lowest achievers however, both genders are equally represented in all cycles."

* Significant gap (p < 0.05)

Table 2 Gender gaps in percent of group membership related to mathematics achievement, grade 8

Percentile	Cycle	Australia	Canada	England	Hong Kong (SAR)	Hungary	Islamic Republic of Iran	Israel	Italy	Japan	Republic of Korea
Upper 20th	1995	-1,0	-0,4	-5,3	-7,9 *	-1,1	-7,5 *	-12,5 *	-3,4 *	-6,3 *	-6,4 *
	1999	-2,4	-2,3	-9,0 *	-1,6	-3,3	-7,9 *	-8,0 *	-4,7 *	-4,1 *	-1,8
	2003	-4,8		0,5	-0,7	-3,8 *	3,2	-6,4 *	-4,1 *	-5,0	-4,2 *
	2007	-7,2 *		-1,8	-1,0	-0,4	1,0	-1,4	-3,7 *	-3,6	-3,3 *
	2011	-3,8		-1,0	-0,6	-2,0	-3,6	-1,1	-5,5 *	-5,2 *	-3,2 *
	2015	-1,1	-4,0 *	3,1	-6,0 *	-2,7	-0,8	-3,8 *	-3,9 *	-0,6	-2,9 *
Lower 20th	1995	-3,1	-0,7	3,3	3,7	-1,8	8,4 *	8,0 *	1,3	0,6	4,3 *
	1999	-1,6	0,7	3,9	-1,2	-0,1	8,2 *	1,4	2,1	0,6	1,7
	2003	3,4		-0,7	-2,3	0,9	-4,3	-0,1	-0,2	-2,2	-0,6
	2007	3,6		1,0	-6,6 *	-0,6	-3,5	-2,8	1,6	-0,5	-0,2
	2011	2,0		-1,9	-4,3 *	2,0	0,1	-5,9 *	3,6	0,8	1,3
	2015	0,5	0,0	0,8	-0,8	3,3 *	-3,2	-1,1	1,7	-2,5	-2,3

Percentile	Cycle	Kuwait	Lithuania	New Zealand	Norway	Russian Federation	Singapore	Slovenia	Sweden	Thailand	United States
Upper 20th	1995	2,0	1,1	-3,6	-1,9	-3,4 *	1,3	-2,7 *	0,5	2,1	-3,7 *
	1999		-2,3	0,7		-2,7	-2,9	-3,1		1,1	-4,7 *
	2003		0,0	-1,8	-0,7	-0,6	0,5	-1,5	-1,5		-3,1 *
	2007	2,9	1,5		-0,2	0,2	1,5	-2,1	0,2	5,2 *	-2,5 *
	2011		1,8	-8,5 *	0,9	-1,1	-0,9	-2,6	2,2	2,5	-1,9 *
	2015	-2,8	-1,5	-0,6	-0,6	-3,1	0,0	-1,2	-3,7	3,1	-1,9
Lower 20th	1995	-11,7 *	-3,5 *	-1,7	-0,2	-3,1 *	0,3	1,8	0,2	-2,3	-0,6
	1999		-0,1	-3,5		-1,5	-0,7	-1,9		-1,0	0,2
	2003		-3,3	-4,2	-2,4	-2,5	-5,8 *	-3,4	-0,7		-0,1
	2007	-12,8 *	-2,9 *		-3,7 *	-2,9	-6,6 *	0,4	-2,5	-9,4 *	1,0
	2011		-4,3 *	4,0 *	-1,0	-0,9	-6,1 *	1,3	-1,4	-9,1 *	0,4
	2015	-7,4 *	-0,1	-2,4	-0,5	3,7 *	-5,3 *	1,1	1,1	-8,6 *	-0,6

Positive gap values and red bars indicate higher relative percentages of girls in the group

* Significant gap (p < 0.05)

than at the upper end of the achievement distribution of this grade. Further, gaps can be observed in both directions, with a higher relative percentage of boys as well as the opposite, a higher percentage of girls.

In 1995, six countries had significant gender differences, with three countries showing relatively more female students in this group, and another three with relatively more boys. Twenty years later, five (but now mostly other) countries still showed significant gender differences, two countries with relatively more girls, and three countries with relatively more boys. No generalizable trend can be observed for this group.

In Hong Kong (SAR), Kuwait, Lithuania, Singapore and Thailand, boys were significantly more likely to be among the low-performing students than girls in more than one cycle. In none of these countries could a tendency towards increased gender equality be observed. While girls were more likely to be in the group of low-performing students in England, Iran, Israel and Korea in the early cycles, this was no longer the case by 1999 or 2003. Russia showed a tendency toward a reversed gender inequality: while boys were overrepresented in this group in earlier cycles of TIMSS, significantly more girls belong to this group in 2015.

Trends in gender gaps in science achievement at grade four
Twenty percent highest performing students (above 80th percentile)

Boys were notably overrepresented in the group of the 20% highest performing students in science achievement at grade four across countries and cycles (see Table 3). This finding is consistent with observations related to mathematics. Similarly, Kuwait was again a remarkable exception, with having constantly more female high-achievers (gap significant in 2007 and 2011).

In 1995, 14 countries had significant gender gaps, all in favor of boys. In 2015, this number reduced to nine countries, indicating some reduction of the gap.

Twenty percent lowest performing students (below 20th percentile)

There were no clear group patterns for the bottom 20% of students in science achievement at grade four (see Table 3). Gender differences were smaller than for the high-achiever group for many countries and while some countries had more boys than girls, others had more girls than boys.

In 1995, eight countries showed significant differences in the relative percentages of male and female students in this group, six with relatively more girls, and two with more boys. In 2015, only three countries out of these eight still showed significant differences, one with a higher percentage of low-performing girls, and two again with relatively more low-performing boys. Kuwait again proved an exception, with remarkably large gender gaps (up to 18% more boys in the group). Differences in all other countries were not significant.

Trends in gender gaps in science achievement at grade eight
Twenty percent highest performing students (above 80th percentile)

Very similar to the findings relating to grade four, the patterns were striking: There were more boys than girls among the 20% highest performing students in science at grade

Table 3 Gender gaps in percent of group membership related to science achievement, grade 4

Percentile	Cycle	Australia	Cyprus	Czech Republic	England	Hong Kong (SAR)	Hungary	Islamic Republic of Iran	Ireland	Japan
Upper 20th	1995	-5,0 *	-4,1 *	-6,9 *	-4,5 *	-6,7 *	-5,6 *	-5,3	-2,9	-4,8 *
	2003	-2,3	-5,1 *		-0,9	-0,3	-3,6	4,8		-3,5 *
	2007	-5,0 *		-6,0 *	-0,4	-3,5	-2,0	0,6		-0,3
	2011	-2,8		-7,9 *	-1,6	-5,5 *	-4,0 *	-2,9	-3,0	-4,2 *
	2015	-1,2	-1,8	-5,8 *	-1,6	-6,0 *	-4,4 *	-0,5	-4,6 *	-4,6 *
Lower 20th	1995	-0,3	1,9	5,1 *	0,1	3,8 *	2,7 *	1,8	-1,4	1,0
	2003	-2,8	1,5		-2,8	-2,4	0,2	-6,1		-0,3
	2007	-1,0		0,4	-2,3	0,3	1,0	-6,0 *		-1,5
	2011	-2,0		4,5 *	-1,7	0,5	-0,4	-1,1	-1,3	1,0
	2015	-1,2	-1,3	1,2	-2,1	2,1	2,5	-5,5	-0,6	-0,4

Percentile	Cycle	Republic of Korea	Kuwait	The Netherlands	New Zealand	Norway	Portugal	Singapore	Slovenia	United States
Upper 20th	1995	-7,5 *	2,9	-9,7 *	-1,2	-5,9 *	-4,4 *	-4,2 *	-4,1 *	-3,9 *
	2003			-4,6 *	0,7	-0,3		-2,7	-1,8	-3,8 *
	2007		10,8 *	-5,3 *	-2,1	-0,3		-2,7	-1,2	-2,6
	2011	-6,6 *	9,3 *	-6,1 *	-2,3	-3,4	-3,6	-3,2 *	-4,6 *	-5,3 *
	2015	-6,4 *	4,4	-1,5	-2,1	-2,2	-3,4 *	-1,1	-6,6 *	-2,4 *
Lower 20th	1995	4,6 *	-13,7 *	5,3 *	-4,8 *	3,3 *	2,3	-0,7	2,4	0,4
	2003			4,5 *	-1,9	-0,8		-1,8	-2,4	0,3
	2007		-17,8 *	5,2 *	-3,8 *	0,4		-2,3	-1,6	0,9
	2011	1,2	-13,7 *	5,3 *	-2,0	0,9	0,4	0,2	0,1	2,6 *
	2015	3,8 *	-9,4 *	-1,7	-3,3 *	-1,7	3,9	-1,2	-0,2	0,4

Positive gap values and red bars indicate higher relative percentages of girls in the group

* Significant gap (p < 0.05)

Table 4 Gender gaps in percent of group membership related to science achievement, grade 8

Percentile	Cycle	Australia	Canada	England	Hong Kong (SAR)	Hungary	Islamic Republic of Iran	Israel	Italy	Japan	Republic of Korea
Upper 20th	1995	-6,2 *	-7,3 *	-8,2 *	-11,4 *	-8,7 *	-11,0 *	-14,0 *	-5,7 *	-8,7 *	-10,6 *
	1999	-8,8 *	-6,4 *	-11,6 *	-9,5 *	-11,1 *	-10,7 *	-8,2 *	-8,1 *	-8,2 *	-7,1 *
	2003	-10,1 *		-6,0 *	-5,7 *	-9,8 *	1,7	-10,3 *	-5,9 *	-6,4 *	-7,2 *
	2007	-9,1 *		-3,8	-4,3	-6,5 *	2,8	-1,1	-4,6 *	-2,9 *	-5,5 *
	2011	-7,3 *	-3,8 *	-1,4	-3,0	-7,8 *	0,3	-1,8	-6,1 *	-5,7 *	-3,5 *
	2015	-2,6		1,2	-8,0 *	-5,3 *	0,6	-1,5	-5,7 *	-2,3	-4,8 *
Lower 20th	1995	0,7	2,6 *	4,0 *	7,0 *	6,5 *	8,0 *	8,8 *	1,7	3,2 *	8,7 *
	1999	2,6	3,5 *	8,4 *	3,6	6,7 *	9,8 *	0,2	3,3	2,4	6,6 *
	2003	6,0 *		3,8	2,4	8,8 *	0,8	3,3	1,4	1,5	3,7 *
	2007	4,4		1,8	-6,1 *	3,2	-5,9 *	-5,3 *	1,6	-0,5	1,0
	2011	3,4	0,4	-3,1	-3,3	5,6 *	-2,9	-6,2 *	5,0 *	1,8	0,6
	2015	1,1		0,3	1,3	5,5 *	-3,8	-3,3 *	1,4	-2,1	-1,7

Percentile	Cycle	Kuwait	Lithuania	New Zealand	Norway	Russian Federation	Singapore	Slovenia	Sweden	Thailand	United States
Upper 20th	1995		-6,0 *	-9,5 *	-9,1 *	-7,3 *	-4,2	-9,9 *	-6,1 *	-0,4	-7,5 *
	1999		-9,4 *	-4,3		-6,9 *	-9,0 *	-7,8 *		-2,1	-9,0 *
	2003	7,7	-3,1	-5,1	-5,4 *	-3,8 *	-3,4	-4,5 *	-4,4 *		-8,2 *
	2007		-1,1		-1,8	-3,4 *	-2,3	-3,4	-0,6	3,1	-6,0 *
	2011	11,0 *	0,7	-9,8 *	-0,7	-3,7 *	-4,8 *	-2,8	-0,4	1,3	-5,7 *
	2015	5,6 *	-1,0	-2,7	-3,3 *	-2,2	-5,1 *	-0,2	-0,9	3,7	-3,7 *
Lower 20th	1995	-20,8 *	3,1 *	4,2 *	2,9	4,3 *	3,9	8,3 *	3,1 *	0,5	2,5 *
	1999		4,9 *	0,3		5,2 *	3,4	2,1		1,0	2,6
	2003	-20,3 *	0,9	1,6	1,2	3,8 *	-2,0	1,1	2,5		4,0 *
	2007		0,2		-2,0	0,7	-4,3 *	-0,6	-2,1	-9,0 *	1,6
	2011	-18,7 *	-5,2 *	4,1	-1,4	1,7	-3,2	-0,3	-2,3	-9,2 *	2,2
	2015		-2,1	-2,6	-0,9	1,7	-2,6	-2,4	-2,0	-10,1 *	-0,2

Positive gap values and red bars indicate higher relative percentages of girls in the group

* Significant gap (p < 0.05)

eight in most countries and cycles (see Table 4). In 19 out of the 20 countries investigated, significant gender gaps for this group were evident in two or more cycles.

In 1995, 17 countries showed a significant gender gap, all in favor of boys. In 2015, only nine countries had a significant gap, with only one country (Kuwait) having more girls than boys in this group. A tendency towards a reduction of the gender gap can be observed in many countries. In England¸ Iran, Israel, Lithuania, Slovenia and Sweden, the gender differences were significant in earlier TIMSS cycles but, by 2015, they were no longer significant. This suggests these countries improved gender equality among their top-performing eighth grade students in science during the last decade.

Twenty percent lowest performing students (below 20th percentile)

Overall, gender inequalities are less pronounced among the 20% lowest performing students compared to the 20% best performing students (see Table 4). The differences, however, mostly favor boys, meaning here that there is, in many cases, a higher percentage of female students in this group compared to the percentage of male students.
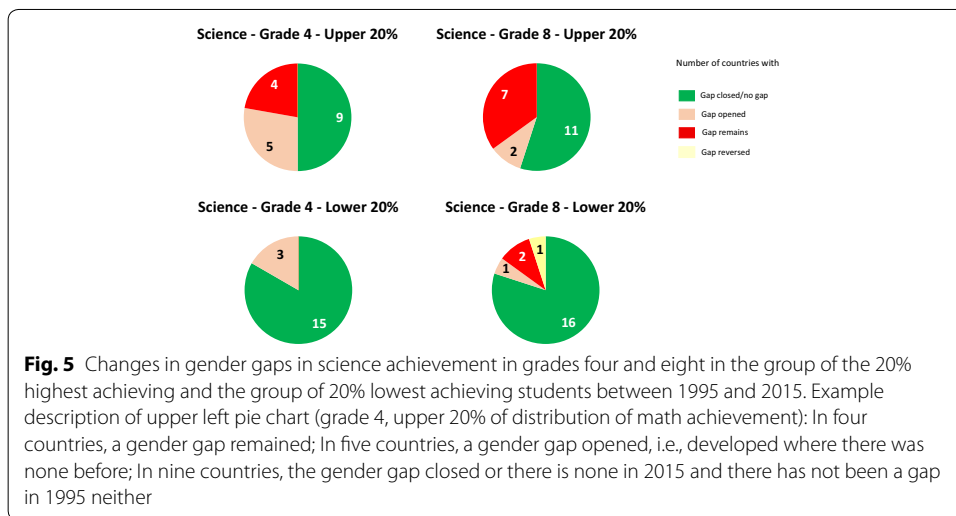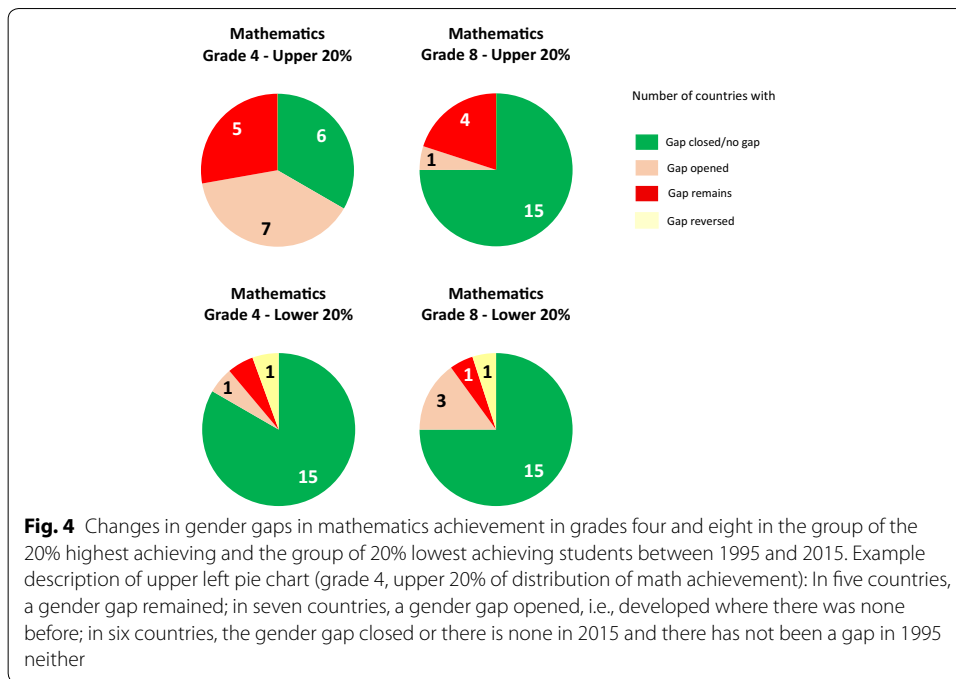
In 1995, 14 countries had significant gender gaps, 13 countries with a relatively higher percentage of girls, and one country (Kuwait) with a relatively higher percentage of boys. In 2015, only four countries had significant gaps, one with a higher percentage of girls, and three with a higher percentage of boys. This trend shows that the gap favoring boys (a lower percentage of boys in the low-performing group) is closing and even beginning to reverse.

Canada, England, Hong Kong (SAR), Japan, Korea, New Zealand, Russia, Slovenia, Sweden and United States have managed to close a previously existing significant gender gap over the last 20 years. Israel has reversed its gender gap: while relatively more girls belonged to the low-performing group in 1995, boys were overrepresented in 2015. A similar (but insignificant) tendency can also be observed in other countries.

A remarkably large gender gap within the group of low-performing students can be observed in Kuwait: three out of four students are male; 30% of all male students in the country are among the 20% of eighth grade students performing lowest in science, while this is the case for only 9% of all female eighth graders. This difference seems to be stable over the last 20 years.

Trends in gender gaps in mathematics achievement across countries

When looking at the trends in gender differences in mathematics achievement across countries, we can observe differences (i) between the two tails of the mathematics achievement distribution and (ii) between grades four and eight. As Fig. 4 shows, in the majority of countries (12 out of 18) there are gender differences at grade four, either persisting or developed newly since 1995, within the group of 20% highest achieving students in mathematics. In six countries, there is no gap: either there has been none in 1995 already, or a gap existing in 1995 closed in 2015. On the contrary, there is no gender gap (any longer) in 2015 in the majority of countries (15 out of 18) for fourth graders at the lower end of the achievement distribution (20% lowest achieving students in mathematics). A similar pattern exists for students at eighth grade. For students at both tails of the mathematics achievement distribution, there are significant gender differences in the majority of countries (15 out of 20).

Fig. 4 Changes in gender gaps in mathematics achievement in grades four and eight in the group of the 20% highest achieving and the group of 20% lowest achieving students between 1995 and 2015. Example description of upper left pie chart (grade 4, upper 20% of distribution of math achievement): In five countries, a gender gap remained; in seven countries, a gender gap opened, i.e., developed where there was none before; in six countries, the gender gap closed or there is none in 2015 and there has not been a gap in 1995 neither

Fig. 5 Changes in gender gaps in science achievement in grades four and eight in the group of the 20% highest achieving and the group of 20% lowest achieving students between 1995 and 2015. Example description of upper left pie chart (grade 4, upper 20% of distribution of math achievement): In four countries, a gender gap remained; In five countries, a gender gap opened, i.e., developed where there was none before; In nine countries, the gender gap closed or there is none in 2015 and there has not been a gap in 1995 neither

Trends in gender gaps in science achievement across countries

Figure 5 shows the change in gender differences in science achievement for the 20% highest and lowest achieving students at fourth and eighth grade in a similar way. While there are some similarities in the overall picture to the findings regarding mathematics achievement, we see also different results. Again, for fourth grade students, differences are found in more countries at the upper tail of the achievement distribution than at the lower end. In half of the countries (9 out of 18), there are persisting or newly developed gender differences for the 20% highest achieving students. For the 20% lowest achieving students, however, gaps have closed or never existed in the majority of countries (15

out of 18). We find rather similar patterns for eighth grade students. Almost half of the countries show gender differences for the 20% highest achieving students (9 out of 20), with mostly persisting gaps, whereas for the 20% lowest achieving students there are no differences in the majority of countries (15 out of 20).

Comparing trends in gender gaps in mathematics and science achievement across countries

Overall, we observe fewer countries with significant gender gaps at the lower tail of the achievement distribution. This holds for both subjects, mathematics and science, as well as for both grade 4 and grade 8. In the upper tail of the achievement distribution, i.e. the 20% highest achieving students, there are more countries with existing gender differences. Here, we find more countries with gender differences for the lower grade students (grade 4) than for the upper grade students (grade 8). However, as the selection of countries included in this analysis is different for grade 4 and grade 8, a direct comparison between results from the two grades is not appropriate.

Trends in gender gaps in mathematics achievement within cohorts

We also examined the gender gaps in mathematics achievement and their trends following up two cohorts from grade four to eight (Table 5).

   While most of the figures and bars displayed in Tables 5 and 6 resemble information from Tables 1, 2, 3, 4 in a different format, the columns "Gap difference between grade 4 and 8" show the development of gender gaps over 4 years of schooling within the same cohort of students. The first cohort represents students attending grade four in 1995 and grade eight in 1999. The second cohort represents students attending grade four in 2011 and grade eight in 2015.

Twenty percent highest performing students (above 80th percentile)

The first cohort (students attending grade four in 1995) exhibited a significant gender gap favoring boys in England, Iran, Japan and Korea (Table 5, left upper part). For the first three of these countries, the gap persisted or even widened by grade eight. In Korea, however, the gap decreased over those 4 years and was no longer significant at the eighth grade. Conversely, a gender gap opened up between grades four and eight in the United States. The gap changes were not significant in any country.

   In two countries (Hong Kong (SAR) and Korea), the second cohort (fourth graders in 2011) contained significantly more male students among the top-performers at both grades. Australia, Hungary, Slovenia and the United States managed to close an existing gender gap in favor of boys at the fourth grade over the ensuing 4 years: the gaps were no longer significant at grade eight. Finally, Slovenia and England showed a significant change in the gender gaps between grades four and eight. In both cases, more boys belonged to the high-performers at grade four. However, while Slovenia achieved gender equality at grade eight, in England, the inequality gap had reversed, and by 2015 favored girls, with significantly more girls achieving high mathematics scores than in 2011 at grade four.

Table 5 Gender gaps in percent of group membership related to mathematics achievement, cohort trends between grades 4 and 8

	Cohort 1			Cohort 2		
	1995 (grade 4)	1999 (grade 8)	Gap difference between grade 4 and 8	2011 (grade 4)	2015 (grade 8)	Gap difference between grade 4 and 8
Upper 20th percentile (high performing students): Positive gap values and red bars stand for higher (or increasing) percentages of girls in the group						
Australia	-2.0	-2.4	-0.4	-3.9 *	-1.1	2.9
England	-4.0 *	-9.0 *	-5.0	-4.1	3.1	7.2 *
Hong Kong, SAR	-2.0	-1.6	0.4	-6.4 *	-6.0 *	0.4
Hungary	-2.1	-3.3	-1.2	-3.7 *	-2.7	1.1
Iran, Islamic Republic of	-6.2 *	-7.9 *	-1.8	-1.9	-0.8	1.0
Japan	-3.9 *	-4.1 *	-0.1	-2.6	-0.6	2.1
Korea, Republic of	-5.0 *	-1.8	3.2	-4.6 *	-2.9 *	1.6
New Zealand	-0.6	0.7	1.3	-2.0	-0.6	1.4
Singapore	-0.2	-2.9	-2.6	0.5	0.0	-0.4
Slovenia	-1.2	-3.1	-1.8	-6.0 *	-1.2	4.8 *
United States	0.0	-4.7 *	-4.7	-5.1 *	-1.9	3.2
Lower 20th percentile (low performing students): Positive gap values and red bars stand for higher (or increasing) percentages of girls in the group						
Australia	0.6	-1.6	-2.2	0.3	0.5	0.2
England	0.6	3.9	3.3	-1.6	0.8	2.4
Hong Kong, SAR	1.3	-1.2	-2.5	1.2	-0.8	-2.0
Hungary	0.0	-0.1	0.0	-0.7	3.3 *	4.0
Iran, Islamic Republic of	4.5 *	8.2 *	3.7	-2.0	-3.2	-1.2
Japan	-0.5	0.6	1.1	0.4	-2.5	-2.9
Korea, Republic of	4.8 *	1.7	-3.1	0.7	-2.3	-3.1
New Zealand	-5.7 *	-3.5	2.2	-1.6	-2.4	-0.8
Singapore	-3.2 *	-0.7	2.5	-2.6 *	-5.3 *	-2.7
Slovenia	1.0	-1.9	-2.9	2.1	1.1	-1.0
United States	-1.2	0.2	1.3	2.0 *	-0.6	-2.6

Interpretation example: In England, in the group of the highest performing students (upper 20th percentile) the proportion of male students was 4% higher than the proportion of female students at grade 4 in 1995 (significant gap). Four years later, this gap has increased to 9% (significant gap). The gap change of 5% however was insignificant

* Significant gap (p < 0.05)

Table 6 Gender gaps in percent of group membership related to science achievement, cohort trends between grades 4 and 8

	Cohort 1			Cohort 2		
	1995 (grade 4)	1999 (grade 8)	Gap difference between grade 4 and 8	2011 (grade 4)	2015 (grade 8)	Gap difference between grade 4 and 8
Upper 20th percentile (high performing students): Positive gap values and red bars stand for higher (or increasing) percentages of girls in the group						
Australia	-5,0 *	-8,8 *	-3,8	-2,8	-2,6	0,2
England	-4,5 *	-11,6 *	-7,1 *	-1,6	1,2	2,8
Hong Kong, SAR	-6,7 *	-9,5 *	-2,8	-5,5 *	-8,0 *	-2,5
Hungary	-5,6 *	-11,1 *	-5,5 *	-4,0 *	-5,3 *	-1,3
Iran, Islamic Republic of	-5,3	-10,7	-5,4	-2,9	0,6	3,5
Japan	-4,8 *	-8,2	-3,4	-4,2 *	-2,3	1,9
Korea, Republic of	-7,5 *	-7,1 *	0,4	-6,6 *	-4,8 *	1,8
New Zealand	-1,2	-4,3	-3,1	-2,3	-2,7	-0,4
Singapore	-4,2 *	-9,0 *	-4,8	-3,2 *	-5,1 *	-1,9
Slovenia	-4,1 *	-7,8 *	-3,6	-4,6	-0,2	4,4
United States	-3,9 *	-9,0 *	-5,1 *	-5,3 *	-3,7 *	1,6
Lower 20th percentile (low performing students): Positive gap values and red bars stand for higher (or increasing) percentages of girls in the group						
Australia	-0,3	2,6	2,8	-2,0	1,1	3,1
England	0,1	8,4 *	8,3 *	-1,7	0,3	2,0
Hong Kong, SAR	3,8 *	3,6	-0,2	0,5	1,3	0,8
Hungary	2,7 *	6,7 *	4,0	-0,4	5,5 *	5,9 *
Iran, Islamic Republic of	1,8	9,8 *	8,0 *	-1,1	-3,8	-2,6
Japan	1,0	2,4	1,4	1,0	-2,1	-3,1
Korea Republic of	4,6 *	6,6 *	2,1	1,2	-1,7	-2,9
New Zealand	-4,8 *	0,3	5,1	-2,0	-2,6	-0,6
Singapore	-0,7	3,4	4,1	0,2	-2,6	-2,8
Slovenia	2,4	2,1	-0,3	0,1	-2,4	-2,5
United States	0,4	2,6	2,1	2,6 *	-0,2	-2,8 *

Interpretation example: In England, in the group of the highest performing students (upper 20th percentile) the proportion of female students was 4.5% higher than the proportion of male students (upper 20th percentile). Four years later, this gap has increased to 11.9% (significant gap). The gap change of 7.1% is significant

* Significant gap (p < 0.05)

Twenty percent lowest performing students (below 20th percentile)

In the countries considered, there was no general trend in terms of the gender composition of the bottom twenty percent of students in mathematics, nor were there identifiable trends over the years or among cohorts (lower part of Table 5).

The picture is very diverse across countries, particularly for the first cohort. Four countries showed significant gender differences in the group allocation; two of them, Iran and Korea, had significantly more girls in this group, while the other two, New Zealand and Singapore, had more boys. These gender gaps reduced over the ensuing 4 years in all countries but Iran, where the gap doubled instead. Moreover, Iran was the only country with a significant gender gap in this group at grade eight in 1999.

Regarding the second cohort, only very few significant gender differences existed in the eleven countries in both grades. In Singapore, significantly more boys belonged to the group of low-performers at fourth grade, and this percentage had doubled by grade eight. A small, yet significant gender difference in the United States with a higher proportion of female students vanished over the years and could no longer be observed at grade eight.

Trends in gender gaps in science achievement between grade four and eight
Twenty percent highest performing students (above 80th percentile)

In contrast to the gender gap trends in mathematics achievement, science achievement showed a very clear and quite generalizable pattern (see upper part of Table 6). Male students were overrepresented in the group of the upper 20th percentile of science achievers at grade four in all considered countries in both cycles (1995 and 2011). This overrepresentation was even more pronounced 4 years later at grade eight, with New Zealand being the only country where this group allocation was insignificant at both grades (cohort 1). Further, there was a significant increase in male students in this group in three countries (England, Hungary and the United States). Fortunately, however, the second considered cohort (students that attended grade four 16 years later) painted a less severe picture. While, similarly, relatively more male students were among the top-performers at grade four, 4 years later this was only still true in five countries (Hong Kong, SAR, Hungary, Korea, Singapore and the United States). The gap did not widen significantly in any country.

Twenty percent lowest performing students (below 20th percentile)

The gender gap is less pronounced in the group of low performing students (Table 6, lower part). In 1995, at grade four only four countries showed significant gender gaps. In Hong Kong (SAR), Hungary and Korea more female students were among the low-performing students, but more male students were in this group in New Zealand. While New Zealand reached gender equity 4 years later, the gap persisted in the other countries and widened in England and Iran, again with a higher proportion of girls.

However, the cohort attending grades four and eight 16 years later showed minimal gender gaps at both grades, with the exception of Hungary, where again more girls comprised the lowest percentile of science achievers at eighth grade. Patterns indicate a tendency towards an increase in male students in this group, a yet insignificant trend that should be closely monitored in the future.

Changes over time

Overall, our findings suggest that girls are catching up with boys. In the group of high achieving students in both subject areas, the overrepresentation of boys continued or even extended from grade four to eight from 1990 to 1995. 16 years later, the overrepresentation of boys found at grade four in 2011 rather reduced at grade eight in 2015. For the lower achieving students, at least regarding science achievement, we can observe a similar change. From grade four in 1995 to grade eight in 1999, there was an increase in the overrepresentation of girls in that group. However, there was no such increase from grade four in 2011 to grade eight in 2015. Even more, the overrepresentation of girls at grade four was much lower already in 2011 than that in 1995.

Discussion, conclusions and policy implications

We investigated the gender gaps in mathematics and science knowledge at both extremes of students' ability distributions. We found that gender gaps that existed 20 years ago have persisted into the present, but also identified encouraging evidence that gender equality in education is increasing. Moreover, data suggests that no general favorable genetic disposition of male students towards mathematics and/or science exists. Otherwise, patterns would be consistent across countries and time. Overwhelmingly obvious is, however, the persisting trend of more male students in the group of high-achievers for both mathematics and science in many educational systems. These subjects have a long history of being more often favored by male students, a situation that fosters gender differences in academic competencies and an underrepresentation of woman in scientific careers. Male and female students may benefit from different teaching approaches and methods to motivate engagement. Several countries have adopted initiatives to address this problem, and the findings indicate that some may have shown success. Similarly, at the lower end of the ability distributions, male and female students are not always equally represented. Patterns differ between countries and cycles. In many countries, male students are overrepresented in this group, while in others, more female students are at risk, predominantly in the upper grades. Policymakers should closely monitor the development of these gaps and initiate measures to tackle gender inequalities. The trends identified in this paper included promising changes in several countries that were able to diminish gender differences in mathematics and science achievement, that have been existing in the past. Furthermore, findings suggest that girls in general are catching up. A closer look at the specific contexts and policy changes might reveal successful measures to counteract gender differences.

This paper adds to existing research into gender gaps in mathematics and science education over the last 20 years, and offers a new approach to gender gap analysis. Investigating the tails of the achievement distributions provide a more differentiated picture of potential gender differences. It thus extends findings of analysis comparing the mean achievement of female and male students. For example, Mullis et al. (2016) report a decrease of the number of countries without achievement differences between boys and girls in math of fourth grade students. However, our analysis revealed that there is quite some variation between high and low achieving students (cf. Tables 1 and 5). For the highest achieving 20% students, the number of countries showing gender differences is rather increasing. For the lowest achieving 20% students, most of the countries included

in this analysis showing no gender differences (any longer). This example indicates that the approach can reveal more (detailed) information on gender differences and their changes over time.

Furthermore, rather than looking at groups of students reaching various benchmarks, we focused on the gender composition of the groups of students comprising the 20th highest and lowest achievement percentiles respectively, for each country. This approach overcomes the problem of only very small samples of students reaching the highest or the lowest benchmark in some countries.

Our research revealed trends in these gaps over 20 years of TIMSS, but it does not explain the mechanisms causing these gaps or any of the underlying factors. Further research is needed to understand these mechanisms better and refine implications and recommendations for policy. IEA contextual data is a valuable research resource to uncover such relations. Although this paper focused only on specific countries and cohorts, it may serve as a template for similar analyses of data from other countries and cohorts that have participated in TIMSS, PIRLS or similar large-scale assessments in education.

Finally, the approach of investigating the relative distribution of a characteristic at the tails of an ability distribution within a country or education system could be used for characteristics other than gender as well. As for gender as such a characteristic, we see that in education systems with several (hundreds of) thousands of students in a certain grade or within a certain age group, this characteristic is fairly equally distributed. That might be very different for other characteristics, for example, students' family background, ethnicity or other student characteristics, for which inequality is perceived as an issue of concern. One of the advantages of the suggested approach is that it is robust in respect to non-equal distributions of such characteristics.

Further research needs

A secondary aim of this paper was to introduce and evaluate a specific approach to identifying (gender) differences in certain outcomes of education (mathematics and science achievement of fourth and eighth grade students). With this more detailed look at tails of ability distributions, the approach could provide information that is more specific and foster the interpretation and explanation of possible inequalities in education. With the analysis presented, trends over time within a set of countries could be identified and support for some common narratives on gender inequalities could be provided, whereas for other narratives we could not find support. In order to investigate possible correlates for changes in or persistence of inequalities, a more detailed look needs to include country, school and classroom contexts, as well as student characteristics. The TIMSS data provides this kind of information and can serve as a valuable source.

and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. Both authors read and approved the final manuscript.

References
Allison, P. D. (1978). Measures of inequality. *American Sociological Review, 43*(6), 865–880.
Baye, A., & Monseur, C. (2016). Gender differences in variability and extreme scores in an international context. *Large-scale Assessments in Education., 4*(1), 541. https://doi.org/10.1186/s40536-015-0015-x.
Benbow, C. P., & Stanley, J. C. (1983). Sex differences in mathematical ability: more facts. *Science, 222*(4627), 1029–1031.
Bergold, S., Wendt, H., Kasper, D., & Steinmayr, R. (2016). Academic competencies: Their interrelatedness and gender differences at their high end. *Journal of Educational Psychology., 109*(3), 439–449. https://doi.org/10.1037/edu0000140.
Costanzo, A., & Desimoni, M. (2017). Beyond the mean estimate: a quantile regression analysis of inequalities in educational outcomes using INVALSI survey data. *Large-Scale Assessments in Education., 5*(1), 14. https://doi.org/10.1186/s40536-017-0048-4.
Davino, C., Furno, M., & Vistocco, D. (2014). *Quantile regression: Theory and applications. Series in probability and statistics*. New York: Wiley.
EGREES. (2005). Equity in European educational systems: A set of indicators. *European Educational Research Journal, 4*(2), 1–151.
Foy, P. (2017). *TIMSS 2015 User Guide for the International Database*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Retrieved from: http://timssandpirls.bc.edu/timss2015/international-database/downloads/T15_UserGuide.pdf.
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.
Hausmann, R., Tyson, L.T., & Zahidi, S. (2009). *The global gender gap report*. Geneva: World Economic Forum. Retrieved February 28, 2018 from https://www.in.gov/icw/files/Global_Gender_Gap_Full_Report_2009.pdf.
Hedges, L. V., & Nowell, A. (1995). Sex differences in mental scores, variability, and numbers of high-scoring individuals. *Science, 269,* 41–45.
Klasen, S. (2002). Low schooling for girls, slower growth for all? Cross-country evidence on the effect of gender inequality in education on economic development. *The World Bank Economic Review., 16*(3), 345–373. https://doi.org/10.1093/wber/lhf004.
Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2016a). *TIMSS 2015 International Results in Mathematics*. Retrieved March 1, 2018 from http://timssandpirls.bc.edu/timss2015/international-results/mathematics/student-achievement/.
Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2016b). *TIMSS 2015 International Results in Science*. Retrieved March 1, 2018 from http://timssandpirls.bc.edu/timss2015/international-results/science/student-achievement/.
Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2016c). *TIMSS Advanced 2015 International Results in Advanced Mathematics and Physics*. Retrieved March 1, 2018 from http://timssandpirls.bc.edu/timss2015/international-results/advanced/.
Mullis V.S, I., O. Martin, M., & Loveless, T. (2016). *20 Years of TIMSS: International Trends in Mathematics and Science Achievement, Curriculum, and Instruction*. Chestnut Hill, MA.: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA).
Nguyen, H.-H. D., & Ryan, A. M. (2008). Does stereotype threat affect test performance of minorities and women? A meta-analysis of experimental evidence. *Journal of Applied Psychology, 93,* 1314–1334. https://doi.org/10.1037/a0012702.
OECD. (2016). *PISA 2015 Results (Volume 1): Excellence and equity in education. PISA*. Paris: OECD Publishing, https://doi.org/10.1787/9789264266490-en.
Retelsdorf, J., Schwartz, K., & Asbrock, F. (2015). "Michael Can't Read!" teachers' gender stereotypes and boys' reading self-concept. *Journal of Educational Psychology., 1,* 1. https://doi.org/10.1037/a0037107.
Schulz, W., Ainley, J., Fraillon, J., Losito, B., Agrusti, G., & Friedman, T. (2017). *Becoming citizens in a changing world. IEA International Civic and Citizenship Education Study 2016 international report*. Cham, Switzerland: Springer.
Strand, S., Deary, I. J., & Smith, P. (2006). Sex differences in cognitive ability test scores: A UK national picture. *British Journal of Educational Psychology, 76*(3), 463–480.
Trueman, C.N. (2015). *Gender and Educational Attainment*. Retrieved February 28, 2018 from https://www.historylearningsite.co.uk/sociology/education-and-sociology/gender-and-educational-attainment/.
UNESCO (2015a). *Education for All Global Monitoring Report 2015: Gender Summary*. Retrieved December 20, 2016 from http://unesdoc.unesco.org/images/0023/002348/234809E.pdf.
UNESCO (2015b). *Education 2030. Incheon Declaration and Framework for Action. Towards Inclusive and Equitable Quality Education and Lifelong Learning for All*. Retrieved September 30, 2016, from http://www.uis.unesco.org/Education/Documents/incheon-framework-for-action-en.pdf.
UNESCO (2017). *Cracking the code: Girls' and Women's Education in Science, Technology, Engineering and Mathematics (STEM)*. Retrieved March 1, 2018 from https://unesdoc.unesco.org/images/0025/002534/253479E.pdf.

UNESCO. (2018). *Handbook of Measuring Equity in Education*. Montreal: UNESCO Institute for Statistics.

United Nations (2018). *Sustainable Development Goals*. Retrieved February 28, 2018, from http://www.un.org/sustainabl edevelopment/sustainable-development-goals/.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.