GENDER DIFFERENCES IN ACHIEVEMENT

A. Authors' Names and Details

Monica Rosén, University of Gothenburg, Gothenburg, Sweden, monica.rosen@ped.gu.se
Isa Steinmann, University of Oslo, Oslo, Norway, isa.steinmann@cemo.uio.no
Inga Wernersson, University West, Trollhättan, Sweden, inga.wernersson@hv.se

Corresponding author: Monica Rosén
Authors are listed alphabetically and have contributed equally to this chapter.

B. Abstract

The relative school achievement of female and male students varies over time and domains. During the last decades, female participation in primary schooling has increased in countries at the low end of the income scale. At the same time, in countries at the high end of the same scale, male underachievement has gained much attention. Gender differences in school achievement cannot be understood purely as a result of classroom activities or of schools as social systems. Factors in society have to play an important role as well, since countries vary drastically in the same assessments of student gender gaps. How men and women are positioned in power structures, in the labor market, and in the family is considered important. Cultural traditions affecting social relations, the understanding of the self, and different attitudes towards schooling are always present. However, school achievement is important for individual life opportunities and achievement differences between females and males are markers of a society's gender structure.

In this chapter, we systematically review research on gender differences across several domains, including mathematics, science, reading, civics, and computer and information literacy. The focus on international large-scale assessment (ILSA) data allows to compare gender gaps across countries or regions, across domains, and over time. In a first step, we systematically reviewed the primary reports of main ILSA studies with respect to gender gaps in achievement. Second, we systematically reviewed 42 secondary analyses of ILSA data that investigated correlates of gender differences in achievement and therefore add a more explanatory perspective to the chapter.

C. Keywords

Female, male, boys, girls, gender differences, international assessment, equality, equity, school age

D. Introduction

When investigating gaps between boys' and girls' achievement in empirical studies, such as the international large-scale assessments (ILSAs), it is important to revisit large societal implications of gender as well as theoretical underpinnings, and finally implications of gender in learning environments such as schools. This introduction will therefore start with embedding school achievement gender gaps in a broader context, before focusing on gender gaps in student achievement in the ILSAs in the main part of the chapter.

When discussing educational gender gaps, it seems worth to keep in mind that formal education (and especially higher school levels) was only open to boys, and to boys of the upper class, for a long time worldwide (e.g., Ariès, 1962). These differences in learning opportunities and school enrollment were explicitly implemented to keep up social boundaries and power structures, including the disparities between men and women. Comprehensive education for all boys and all girls is a rather new phenomenon, especially in lower-income countries. With the shift to comprehensive education driven by democratic, and feminist movements, the ideal of equal education opportunities for boys and girls becomes more and more established. In most parts of the world, gender gaps in education are nowadays actively tackled. The United Nations included the goal of equal education for boys and girls as one of the core sustainable development goals (e.g., UNESCO, 2019).

Historically, however, empirical research on gender gaps in school achievement and underlying cognitive abilities was intended to proof a 'male superiority' that was assumed to exist and that would justify gender differences in educational opportunities and other areas of the social and occupational life (e.g., Shields, 1982). One of the first to question if the common gender roles were determined by nature was the anthropologist Margret Mead (1935). She found cultures in Papua New Guinea with reversed gender patterns compared to what was viewed as natural in the Western world. Also, when reviewing and meta-analyzing empirical findings (e.g., Halpern, 2012; Maccoby & Jacklin, 1974; Willingham & Cole, 1997), it turned out that boys and girls performed very similar on cognitive tasks in standardized tests. In fact, the only robust cognitive gender differences that were identified were small advantages of boys in visual-spatial and mathematical tasks, and small advantages of girls in verbal tasks. In most cognitive domains, boys and girls performed very similar. Rosén (1998) further found that girls performed better at a general level and males at narrower specializations. Larger gaps were found in non-cognitive areas (e.g., throwing distance, aggressiveness). Therefore, these studies did by no means support hypotheses about big differences in the functioning of the male and female brain. On the contrary, Hyde (2005) concluded in her meta-analysis of meta-analyses that males' and females' psychological characteristics are mostly alike, which she summarizes as the gender similarity hypothesis.

Although women and men have therefore very similar cognitive prerequisites for education, training, and labor market participation, we can still observe partly large gender disparities in many areas of life and across countries (e.g., Saini, 2017). Beliefs about large gender differences and social roles of boys and girls and men and women still prevail. For instance, women still earn much less than men in almost all countries in the world and men are still less involved in taking care of their children (UNDP, 2019; World Economic Forum, 2019). Differences between boys and girls, related to the individual body, and gender differences, related to social constructions and individual minds, are still controversial issues, politically as well as scholarly. When looking at the education sector, we see a global increase in the school enrollment and attainment of both boys and girls (UNESCO, 2019). In lower income areas, girls have however still not the same probability to go to primary and especially secondary schools, and to finish education levels with formal certificates. In high-income countries, the enrollment rates of boys and girls are nowadays typically close to 100 % for both boys and girls. Usually, girls have even better chances to attend secondary schools, than boys in these countries (ibid.). Also, girls are usually found to on average attain better grades and higher school leaving certificates (Arnot et al., 1999; UNESCO, 2019; Wernersson, 2010).

Many theories have been proposed to explain why gender matters in so many areas of life, including school achievement (e.g., Connell, 2002; Halpern, 2012; Maccoby, 1998). These theories emphasize different societal and cultural factors and actors. It is however clear that gender gaps in school achievement cannot (only) be explained by innate differences in cognitive abilities, because the gender gaps in cognitive abilities are too small to explain these larger differences, and because the school achievement gender gaps vary so much between countries. Halpern (2012) also stressed the value of comparative studies for gender-related questions:

> "The underlying rationale for cross-cultural research in this area is that all females, everywhere in the world, share the same biology, as do all males. Although there are obvious differences among people in their skin color, hair texture and curl, shape of eyes, etc. the biology of femaleness and maleness is the same everywhere. Except for medical anomalies, members of each sex have the same chromosome configuration for determining sex internal reproductive organ, gonads (sex glands), genitals, and sex hormone balance, although it remains possible that there are groups where environmental variables altered these biological indicators of sex. In addition, the demand for cognitive abilities is universal in that everyone must learn, use information, make decisions, represent and communicate meaning, navigate through space, reason with quantities, create, decide, solve problems, and so on." (p. 344).

The present chapter focuses on gender gaps in student achievement, which can be considered one central domain of gender inequalities in education. Student achievement tests measure learning outcomes in a standardized way and can therefore be used to depict gender differences in the actual learning processes.

E. Main Text

Studying Gender Differences in Achievement in ILSAs

As mentioned in the introduction of this chapter, ILSAs provide unique possibilities to investigate gender differences in achievement both in relation to theories about equity and equal opportunities, but also in relation to theories about nature and nurture. If gender differences in achievement would be innate to any degree, or determined by biological factors, a consistent pattern of achievement differences across countries (in which boys and girls attend the same schools) would be expected (e.g., Halpern, 2012; Maccoby & Jacklin, 1974). As we will show below, no such patterns can be seen in any of the ILSAs. The fact that gender gaps vary in size and direction, across age groups, time points, and countries should be regarded as strong evidence for the opposite, that there are no gender differences in the cognitive capabilities. Plausible causes for gender differences in educational achievement are bound to be found elsewhere, for instance, in the cultural settings where they have emerged.

Gender differences in educational achievement are, when observed, usually very small compared to the general inter-individual variation in each measure, and rarely account for more than 2 % of the variation (e.g. Hyde, 2005, 2014; Willingham & Cole, 1997). This typical pattern is illustrated in Figure 1, where the overlap between the two distributions is almost complete. The figure shows normal distributions of two groups with a mean difference of 0.2 standard deviations (SD); in the case of gender gaps, a mean difference of 0.2 SD is not among the smallest between boys and girls. The fact that gender differences are usually small does however not make them unimportant or uninteresting. On the contrary, they have to be interpreted in the context of other gender differences, like the still pronounced gender gaps in the possessions of societal goods and power (e.g., World Economic Forum, 2019).

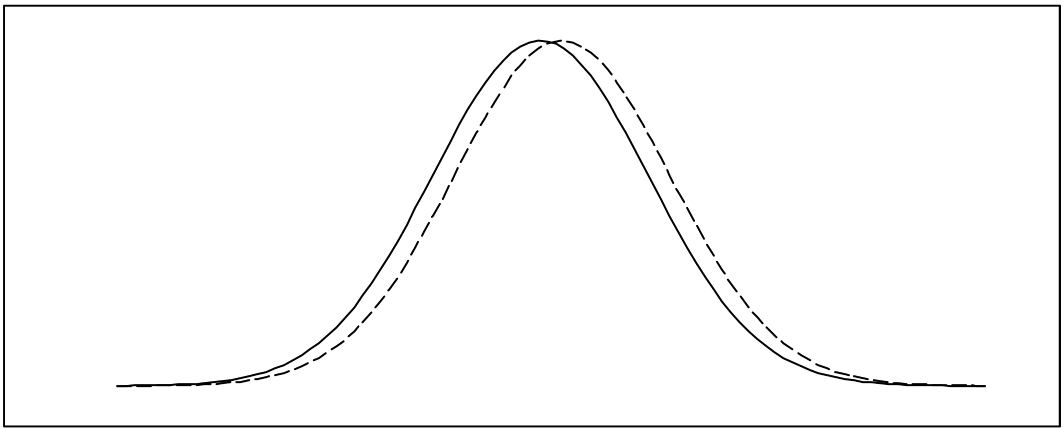

Figure 1. Normal distributions of achievement (x-axis) of two groups (solid and dotted lines) with a mean difference of 0.2 standard deviations (SD). Density on y-axis.

It should also be remembered that gender gaps in achievement are not the only relevant gender gaps in education (e.g., UNESCO, 2019). There are many education-related aspects in which patterns of gender may be both important and pronounced (e.g., grades, attainment, aspirations, friendships, attitudes, stress, wellbeing, course choice, absenteeism). In these respects, the ILSAs provide unique possibilities to investigate gender gaps and gender patterns in relationships between various educational factors and compare cultural differences across countries, as well. This chapter focuses, however, on gender gaps in student achievement.

There are many well-known challenges inherited in the study of gender differences. One that often seems difficult to avoid is the polarization that follows when females are contrasted with males. In that vein follows also a tendency to, albeit incorrectly, generalize group differences to each and every one within that group. As illustrated in Figure 1 for instance, a small group difference cannot be generalized to all individuals in those groups. When randomly looking at one individual from each group, their individual scores can both well be very low or very high. A good solution for how to handle differences and similarities at the same time is still missing. In the study of gender differences, it is therefore important to also highlight similarities, and especially when various cognitive abilities and achievement measures are in focus (e.g., Halpern, 2012).

Research Questions and Methods of Inquiry

The present chapter had two main goals: At first, we aimed to describe both broad trends and country differences in gender gaps in different domains of student achievement. In the first part, we overviewed and summarized findings from official primary reports published by the primary research teams of the ILSAs. We focused on ILSAs of students in compulsory education. Specifically, we focused on (1) the Programme for International Student Assessment (PISA; conducted in three-year cycles between 2000 and 2018), (2) the Progress in International Reading Literacy Study (PIRLS; five-year cycles between 2001 and 2016), (3) the Third International Mathematics and Science Study (TIMSS 1995) respectively the Trends in International Mathematics and Science Study (TIMSS; four-year cycles between 1999 and 2015), (4) the Civic Education Study (CIVED; conducted between 1996 and 2000) and its successor International Civic and Citizenship Education Study (ICCS; in 2009 and 2016), as well as (5) the International Computer and Information Literacy Study (ICILS; in 2013 and 2018). The Organisation for Economic Co-operation and Development (OECD) conducts the PISA study, and the International Association for the Evaluation of Educational Achievement (IEA) conducts the other studies of interest.

The second research aim was to investigate how the data from the mentioned ILSAs were used in secondary research to investigate gender differences in achievement and their relationships to other variables. We therefore conducted a systematic literature review and summary of these findings. In the systematic literature review, we applied the following three inclusion criteria: First, we only included empirical studies that used data from at least one country from PISA, PIRLS, TIMSS, CIVED, ICCS, and/or ICILS. Second, we only included studies that explicitly focused on gender differences in achievement and their relationships with other variables. We therefore excluded studies that used gender only as control variable, only described trends in gender gaps or international differences, investigated gender gaps in achievement and their relationship to method effects (e.g. differences in gaps depending on item format or subdomain), or investigated gender gaps in achievement measures not assessed in standardized tests (e.g., school grades). Third, we focused on studies that were published in English in peer-reviewed journals and in sufficient detail to interpret the coefficients of interest. The search and selection of references was conducted in four steps. At first, we searched for articles in ten literature repositories on 06 December 2019 (see Appendix for further details). This search resulted in 816 references. By using the Citavi software, we removed duplicates and identified 585 references. Second, we screened the 585 references' titles and abstracts for a possible fit to the inclusion criteria. After excluding studies that clearly did not fit to our inclusion criteria, 99 references were identified. Third, we conducted an in-depth review of these full texts regarding their fit to the above-mentioned inclusion criteria and identified 36 studies. Fourth, we conducted an additional manual literature search by screening the publication lists on the IEA and OECD homepages and the references lists of the most recent articles that were identified in the previous step. We identified six additional studies. Therefore, overall 42 studies were included in the literature summary below.

Describing Gender Differences in ILSAs

In the following, we review the gender gap findings that are presented in the official primary reports of the IEA and OECD. In order to increase the readability of the findings, we present all group differences in units of the international standard deviations (SD), which are 100. We group our summary by subject-domains, although some studies assessed more than one subject domain at once (i.e. TIMSS and PISA).

Interestingly, the primary report of the first PISA study in 2000 (OECD, 2001) underscored the closing gender gap in educational *participation*. It is emphasized that twice as many women aged 25 to 34 compared to those aged 55 to 64 had at that time completed tertiary education in 13 of the OECD countries. In 2001, the women's graduation rate was equal or exceeded that of men in many (17 of 30) OECD countries and in all but one of the participating non-OECD countries (OECD, 2003). Gender differences in academic achievement are, however, pointed out as an area that needs to "receive close attention" (OECD, 2001, p. 122) since there seems to be a risk for male under-achievement. More boys (14 %) than girls (9 %) did not reach the PISA baseline level of proficiency in any of the three subject areas measured (reading, mathematics, and science). Another reason for the OECD to focus on gender is the female underrepresentation in traditionally male jobs demanding mathematical and science skills (OECD, 2016). It could be noted that the underrepresentation of men in traditionally female jobs is as pronounced or even larger, which, for unclear reasons, is not described as a problem of the same magnitude.

Gender Gaps in Reading

Reading is a core competence that affects the performance in most or all other school subjects as well as numerous aspects of work, societal participation, and wellbeing in adult life. Reading is not just a simple decoding of signs, but a complex task that involves many cognitive abilities. It is difficult to distinguish reading skills in a technical sense from more general cognitive abilities to interpret, understand, and reason. As described below, reading is also the domain that shows the largest gender differences of those academic achievement domains that are measured in the ILSAs. Reading is assessed in 4th-grade students in PIRLS and in 15-year-old students in PISA in recurring cycles.

Gender Gaps in Reading in Fourth-Grade Students

The first PIRLS was launched by the IEA in 2001. The study assessment frameworks include definitions of central concepts along with descriptions of test and sampling designs and are updated for each study cycle (e.g., Mullis & Martin, 2015). Common for all cycles of PIRLS so far is that the target grade for the assessment is grade 4, or the grade with most 10-year-olds. The study is designed to reoccur every five years. So far, four cycles have been conducted: 2001, 2006, 2011, and 2016. In any educational system, with four years of schooling, most pupils have acquired basic reading skills.

The overall reading literacy scale in PIRLS includes both informational and literary texts, representing different purposes of reading, to gain information and to experience emotions and develop knowledge. The test items in PIRLS are based on four reading comprehension processes: 1) Locating and retrieving, 2) straightforward inferencing, 3) interpreting and integrating and 4) evaluating. Results have been reported on four subscales, two based on reading purposes (informational and literary reading), and two based on levels of cognitive reading processes: 1) Retrieving respectively straightforward inferencing and 2) interpreting, integrating, or evaluating.

With the gradual digitalization of societies, the PIRLS assessment of reading literacy has been extended to include digital reading and authentic online reading tasks. In 2016, countries could choose whether they also wanted to participate in the assessment of online informational reading, called ePIRLS, an optional extension of the regular PIRLS study. In ePIRLS, a simulated internet environment was used to present the students with school-like assignments involving science and social studies topics. Fourteen countries and two benchmarking entities (i.e., subpopulations of countries) participated in ePIRLS 2016.

The number of countries or education systems that participated in PIRLS has increased over the cycles. The latest study from 2016 included 50 countries (Mullis et al., 2017b). Grade 4 girls showed a higher average reading achievement level than boys in all countries except Macao SAR (Special Administrative Region of the People's Republic of China) and Portugal, where the achievement levels were similar (see Figure 2). The score point difference between boys and girls can be interpreted in relation to the SD of the international scale, which is 100. For orientation: The average gain from one school year to the next is about 0.40 SD. The average advantage for girls was 0.19 SD across the 50 countries in PIRLS 2016. Small but significant differences (about 0.10 SD) were observed in 25 countries, and medium differences (about 0.20 SD) in 13 countries. Comparably large differences were found in 10 countries in Africa or the Middle East, where girls outperformed boys with between 0.30 SD to 0.70 SD. Furthermore distinctive for these countries was that they all performed well below the international reading achievement average. The international variation in SD differences between boys and girls can be contextualized against the overall between-country differences in reading achievement. The difference between the highest average reading performance (in the Russian Federation) and the lowest average performance (in South Africa) amounted to 2.61 SD in PIRLS 2016.

The PIRLS 2016 international report also conveyed trend information on gender differences from all previous PIRLS cycles. Twenty-seven countries and benchmarking systems participated in all four cycles, and another fifteen participated in three cycles. There are sparse signs of closing gaps between boys and girls. This general pattern also holds when looking at the trends of gender differences in the reading purpose or cognitive process subscales (Mullis et al., 2004; Mullis et al., 2007; Mullis et al., 2012; Mullis et al., 2017b). In no country or benchmarking entity did boys achieve higher average reading scores than girls.

It is often expected that information technology is a field that implies advantages for boys. For most countries, the gender gap in ePIRLS was, however, similar to the overall reading scale. The average advantage for girls in digital reading was 0.12 SD across the 14 countries that participated in ePIRLS. The average gender gap ranged between 0.02 and 0.29 SD. Girls average achievement was higher everywhere but in Italy, Portugal, and Denmark, where the achievement level was similar for boys and girls. Just as in PIRLS, the findings for the cognitive process subscales were similar in ePIRLS, as well. Mullis et al. (2017a) suggested that technology-related advantages, often enjoyed by boys, may not apply to reading and learning with online information. In the student questionnaire, about 80 % or more of both girls and boys reported that they liked working with the e-tasks.

A general conclusion from this review of fourth-grade reading is that girls perform on average higher than boys on reading tasks in all educational systems (although the difference is not always significant) regardless of the presentation modes of the texts and regardless of the kind of reading tasks. This was also true for online reading. The review of gender differences on the various reading subscales indicated only minor deviations compared with the gender patterns on the overall reading scale.

Gender Gaps in Reading in 15-Year-Old Students

In PISA, the reading skills of 15-year-old students have been assessed every third year since 2000 with the latest in 2018. In PISA, one of the three subject domains reading, mathematics, and science are specifically targeted with more test items and a more extended analysis every third year. Reading was the main subject in 2000, 2009, and 2018. The PISA reading scale is also based on reading of different text types from a wide range of situations and items that require similar cognitive processes as those in PIRLS. According to the PISA framework, which is also updated every new cycle, reading tasks include finding, selecting, interpreting, integrating, and evaluating information from the full range of texts associated with situations that reach beyond the classroom (e.g., OECD, 2019a). As in PIRLS, the reading achievement scale in PISA is also created from a matrix design of reading tasks, where the international mean is set to 500 with an international SD of 100.

The general pattern is the same for 15-year old's as for 10-year old's in PIRLS, with girls performing better in all participating countries in all PISA studies. In 2000, 28 OECD countries participated in PISA. Furthermore, four non-OECD countries participated in 2000 and another 11 non-OECD countries in 2001. Translated into SD units, the overall gender gap across countries was 0.32 SD (OECD, 2002, 2003). The advantage in favor of girls varied between countries from 0.07 SD (in Peru) to 0.58 SD (in Albania). This could be compared to the huge between-country variation in the average performance of all students (boys *and* girls), where the difference between the highest average reading level (in Finland) and the lowest (in Peru) was 2.19 SD in PISA 2000.

Girls performed better on all the reading subscales with the differences tending to be larger for more complex skills (e.g., 0.45 SD on the reflection and evaluation scale). This is thought to be related to different reading habits and attitudes of boys and girls. Girls read more (especially more fiction), they read more difficult texts, and enjoy reading more (OECD, 2001).

In PISA 2009, the next assessment where reading was focused, the overall gender differences were similar to 2000. Across OECD countries, the average gender difference was 0.39 SD, which is described as of the same size as the progress in one year of schooling (OECD, 2010b). In the PISA 2009 assessment, 65 countries or education systems participated and girls significantly outperformed boys in all of them. The gap varied from over 0.50 SD in 14 countries to less than 0.25 SD in 7 countries. In several high performing countries, like Finland and several Asian countries, the differences in favor of girls are large to medium.

Gender differences were most pronounced at the low end of the scale. In OECD countries, one girl of eight and one boy of four does not reach the performance level 2, which is described as a lack of "essential skills needed to participate effectively and productively in society" (OECD, 2010a, p. 12). The size of the gender gap also varied between the subscales in the 2009 assessment. Compared to the combined scale, the access and retrieve subscale showed a similar difference (0.40 SD), the interpreting subscale a somewhat smaller difference (0.36 SD), and the reflecting subscale a larger difference (0.44 SD).

In the most recent PISA assessment in 2018, the size and direction of the gender differences in reading were, with 0.30 SD in favor of girls across countries, very similar to the earlier assessments (see Figure 2). The general level of reading performance in a country tended, however, to be somewhat smaller than the OECD average in countries with a high-level performance (OECD, 2019b).

When the 2018 results are compared with the 2009 results, the gender gap was smaller in most countries (36 out of 64). This was in some countries (17) due to an improved performance of boys, in some (11) due to a decline in girls' performance and in some (5) there was a better general performance *and* a narrower gap (OECD, 2019b).

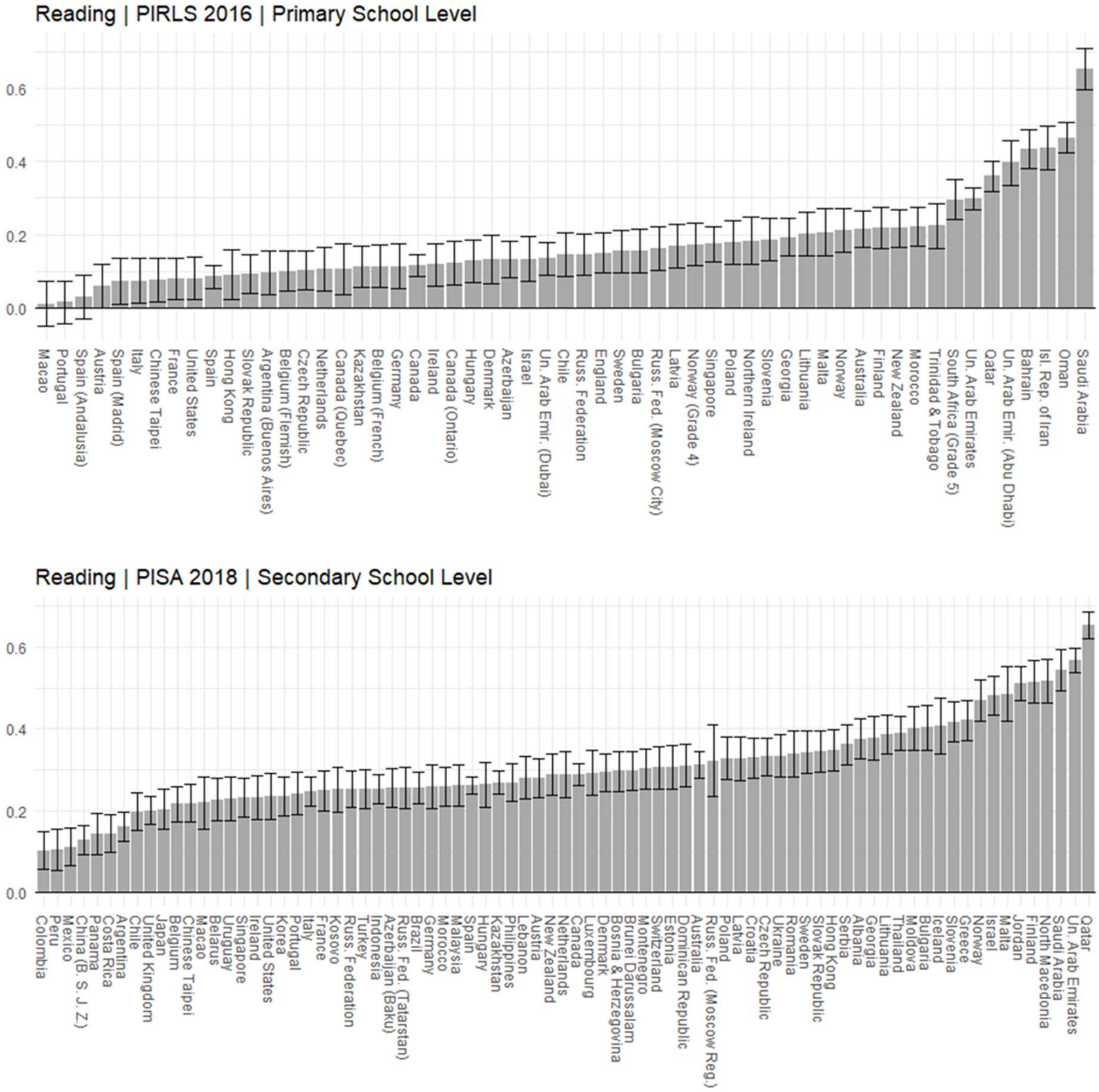

Figure 2. Gender gaps in reading in units of the international standard deviation (SD; y-axis) in all participating countries and benchmarking participants in two recent ILSAs. Values above zero indicate higher scores of girls compared to boys. Larger confidence intervals indicate smaller sample sizes.

Gender Gaps in Mathematics

In this section, we summarize gender differences in mathematics achievement as observed and reported by the IEA in the TIMSS study, and by the OECD in the PISA study. TIMSS has been reporting gender gaps in mathematics (and science) for fourth- and eighth-grade students regularly every fourth year since 1995 (except in TIMSS 1999, when only grade 8 was assessed). So far, TIMSS contains seven measurement points for grade 8 and six for grade 4 with comparable measures of gender gaps in achievement. In addition, the OECD PISA study has reported gender differences in mathematics and science achievement among 15-year-old students every third year since 2000.

With respect to measuring mathematics achievement, both TIMSS and PISA aim to assess school systems in terms of student achievement. However, there are three major differences between the TIMSS and PISA studies. The first is that the IEA tests of mathematics achievement are based on the common school curriculum across participating countries and the test items therefore reflect typical school tasks, whilst the OECD tests in PISA aim to reflect future market or societal expectations. This means that the two studies use indicators from different contexts, and thus, the measured outcome is not identical. Both studies target, however, to a large degree the same underlying skills. The second difference is the definition of target populations, where the IEA selects national samples based on years of schooling and the OECD bases their samples on age (i.e., 15-year-old students). The third difference is the test design, where the coverage of the assessed constructs varies. In TIMSS, the mathematics and science constructs are well-covered in every cycle. In PISA, only one of the three achievement domains mathematics, science, and reading are well-covered in each cycle. Mathematics was the main subject in PISA 2003 and 2012.

In TIMSS 2015 (Mullis, Martin, Foy, & Hooper, 2016), 57 countries participated (49 countries in the grade 4 study, and 39 in the grade 8 study). About 40 of the 57 countries were high-income countries, 10 belonged to the upper-middle income group, and two were lower-middle income countries. In addition, achievement trends were reported for the 43 countries which had participated in two or more TIMSS studies since 1995 (Mullis, Martin, & Loveless, 2016).

Gender Gaps in Mathematics in Fourth-Grade Students

The overall mathematics achievement measure used for grade 4 covers three mathematical content domains (numbers, geometric shapes, and measures and data display), and within each of those three cognitive domains (knowing, applying, and reasoning). These subdomains have been used in all cycles since 1995.

Just as in PIRLS and PISA, the mathematics achievement scales are computed with a mean of 500 score points and an international SD of 100 score points. In TIMSS grade 4, a school year conforms to an approximate difference of 0.40 to 0.50 SD. Across the 49 countries that participated in the grade 4 assessment of TIMSS in 2015, a significant gender gap was only found in 19 countries (Mullis, Martin, Foy, & Hooper, 2016). As can be seen in Figure 3 below, the gaps were typically larger in the countries where the girls achieved the higher mean scores (Finland, Indonesia, South Africa, Oman, Saudi Arabia, Jordan, Kuwait, and Bahrain). The largest gap was reported in Saudi Arabia with 0.43 SD. In 18 countries, boys achieved a higher average than girls, typically around 0.07 to 0.08 SD. Italy showed the largest advantage of boys, with a mean difference of 0.20 SD. In the remaining 20 countries, there was no significant gender difference in grade 4.

Both high- and low-performing countries report gender differences in both directions. Also notable is that the countries that show a female advantage were mainly Middle Eastern countries (except Finland and South Africa), whilst it is mainly South European countries that had a male advantage. The countries with no gender gap in 2015 were located all over the world (Mullis, Martin, Foy, & Hooper, 2016).

With regard to the question about *change* in gender gaps in mathematics achievement in 4th-graders, the TIMSS study offers comparable achievement measures at five time points; 1995, 2003, 2007, 2011 and 2015. There were 17 countries which participated in all five cycles. Reviewing the trends of gender differences revealed a small but stable gender gap, on average 0.05 SD in about half of these countries. Most of the countries with data from a shorter period reported no or decreasing differences regardless of the direction of the gender gap at their first measurement. Increasing gaps are observed in only three countries (Finland, Bahrain, and Saudi Arabia), all three to the female advantage.

In summary, the main observation from TIMSS 2015 and previous cycles was that the gender gap in grade 4 has either decreased or in some cases turned towards a female advantage (Mullis, Martin, & Loveless, 2016). The new release of TIMSS 2019 results revealed no notable changes in grade 4 gender gaps (Mullis et al., 2020).

Gender Gaps in Mathematics in Eighth-Grade Students

In TIMSS 2015 grade 8, the overall mathematic achievement measure covered four sub-domains (number, algebra, geometry, and data and chance) and within each of them also the same three cognitive domains (knowing, applying, and reasoning) as in grade 4 (Mullis & Martin, 2013). For grade 8 mathematics, these subdomains have been part of the study throughout all previous cycles.

Gender gaps in mathematics were found in 13 of the 39 participating countries in the grade 8 population. A higher female average was found in seven countries with a mean advantage of girls of 0.17 SD. Six other countries reported higher male averages varying between 0.04 and 0.18 SD. One may note that most countries that reported male advantages in 2015 were OECD countries, while most countries reporting female advantages were Middle Eastern countries. For 26 countries from different parts of the world, no gender differences were found in grade 8 mathematics achievement at the average level (Mullis, Martin, Foy, & Hooper, 2016).

Changes in gender gaps in grade 8 mathematics achievement has been monitored in TIMSS across six cycles between 1995 and 2015. TIMSS 1999 was the trend field trial, designed to provide comparable measures across time for the very first time. There are 16 countries that participated in all six cycles, 11 of them being OECD countries (Mullis, Martin, & Loveless, 2016). TIMSS 1995 revealed a statistically significant advantage for boys in four of the countries, and in 12 countries there was no statistically significant gender gap (Mullis et al., 2000). Twenty years later, the difference to the boys' advantage has disappeared in all four countries but in another three of the 16 countries a small but statistically significant gap emerged instead (Mullis, Martin, & Loveless, 2016). Reviewing changes for countries with trend data from shorter periods reveals a decreasing female advantage in some countries, decreasing male advantage in only one, and increasing or stable female advantages in some other countries. Mullis, Martin, and Loveless (2016) conclude that the gender gap in mathematics, which in 1995 was usually to the boys' advantage, has in both grade 4 and grade

8 in many countries either decreased, disappeared or in some cases turned towards a female advantage. However, a narrower review indicates that such a turn has only been observed in Finland in grade 4. A female advantage in mathematics seems to be more common in countries that joined the TIMSS study later, and mainly in countries from the Middle East or Africa.

It is nevertheless interesting to note that most countries that monitored their mathematics performance over a long time seem to have succeeded to reduce or remove gender differences in mathematics. Whether this has been accomplished through governmental efforts or not is unclear, but the monitoring of trends indicates that it may take a long time for systems to succeed in these matters, and also that in some countries small differences seem to appear and disappear more or less by chance across time points.

Gender Gaps in Mathematics in 15-Year-Old Students

Across participating countries in PISA 2003, boys tended to score on average higher in mathematics than girls (0.11 SD). In half of the participating 26 countries, there was an advantage for boys that reached statistical significance (OECD, 2004). If there were male advantages in mathematics in the PISA assessment, they were found to be driven by a larger proportion of boys in the relatively small fraction of students performing at the highest levels.

The gender differences in different parts of the distribution followed the same pattern in the PISA assessment in 2012 with more boys among the high performers (OECD, 2015). Boys performed on average somewhat better than girls (international gap 0.08 SD). The largest differences to the boys' advantage were found in some Latin American and European countries. At the same time, girls performed better in nine countries, mostly in Asia and Eastern Europe.

In PISA 2018, boys had a small advantage of 0.05 SD across the participating countries (OECD, 2019b). However, only in 32 of 79 countries, boys performed better than girls and in 14 countries, girls performed significantly better than boys. In most of the countries that participated both in 2012 and 2018 (43 of 64), the gap in mathematics remained the same. In five countries, the difference narrowed because of girls' better performance, but in seven countries the gap narrowed because of a declined performance of boys. The picture is however complex in mathematics performance with varying differences at both ends of the performance distributions: "For example, in many countries girls' scores at the first decile of the distribution of mathematics performance were higher than boys' scores, meaning that the lowest-performing girls scored above the lowest-performing boys in their countries." (OECD, 2019b, p. 149). Among the high-performers, however, boys tended to be overrepresented, but this pattern is not present in all education systems. There are 35 countries and economies (including high-performing systems), where no significant gender gap among the top-performers was found. In Figure 3, we depicted the mathematics gender gaps across countries in mathematics in PISA 2018.

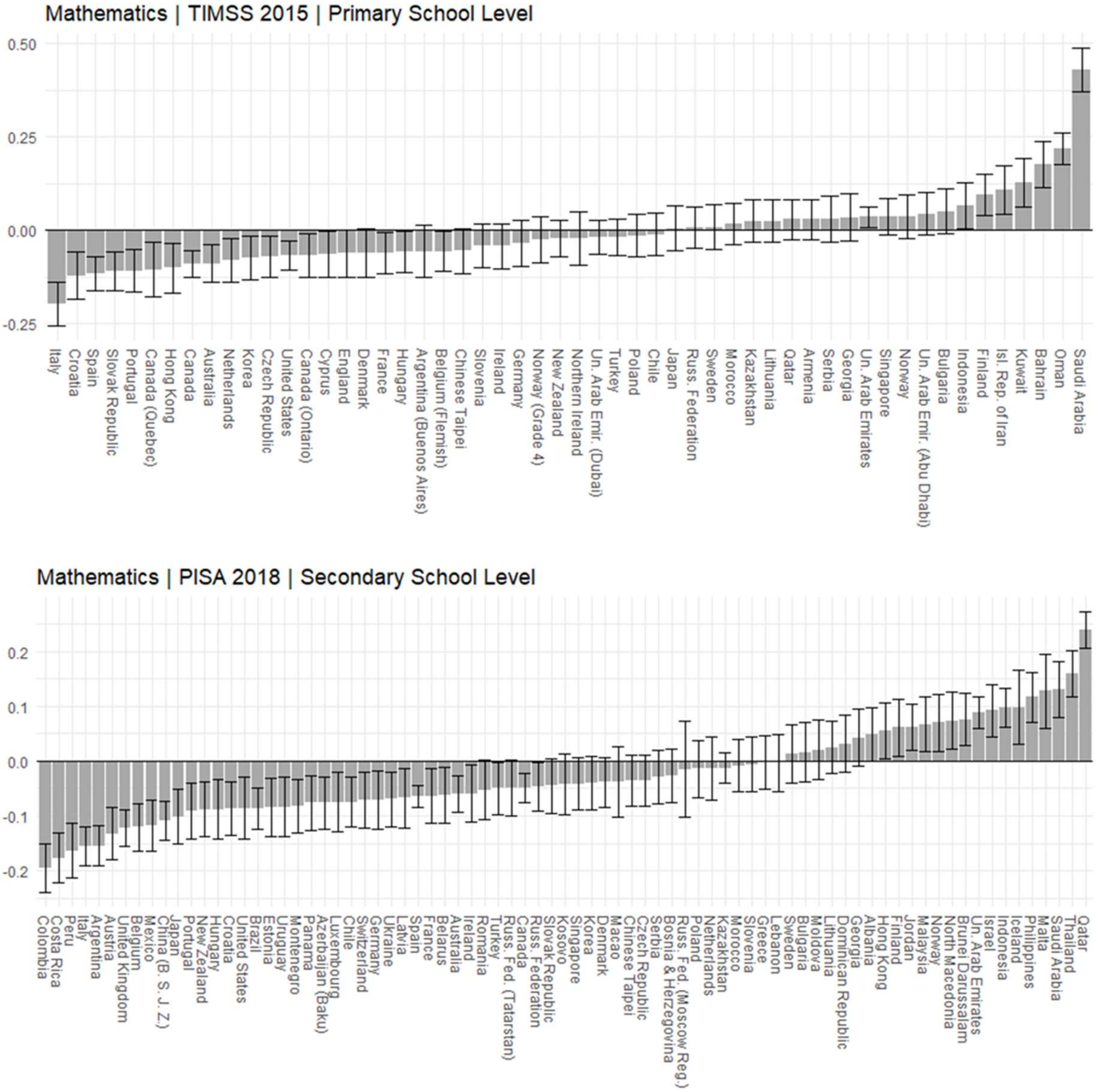

Figure 3. Gender gaps in mathematics in units of the international standard deviation (SD; y-axis) in all participating countries and benchmarking participants in two recent ILSAs. Values above zero indicate higher scores of girls compared to boys. Larger confidence intervals indicate smaller sample sizes.

Gender Gaps in Science

Similar to the reporting of mathematics achievement, the IEA has been reporting gender differences in science achievement in primary and secondary school levels in all their science assessments. Before 1995, two studies of science achievement were conducted, one in 1970-1971, where science was one of the subject domains of the Six Subject Study (Comber & Keeves, 1973), and a second study in 1984, the Second International Science Study (SISS; Postlethwaite & Wiley, 1992). However, starting with TIMSS 1995, the study design was adjusted in two notable ways. At first, the target population was based on students' grade instead of age, and second, the science achievement scale was designed to measure trends both within and between countries. Up until now, seven eighth-grade (1995, 1999, 2003, 2007, 2011, 2015, and 2019), and six fourth-grade measurement points (all except 1999) have been published.

The OECD PISA study has reported gender differences in science achievement since the first in 2000 in three-year cycles, resulting in seven measurement points so far. Science was the main subject of assessment in PISA 2006, and again in 2015. PISA includes science items in every new study cycle, but the achievement scores are based on many more items when science is in focus. Trend measures of science achievement and gender gaps have also been reported for the PISA study among 15-year-old students on a regular basis.

As mentioned in the reading and mathematics sections above, the IEA and OECD studies are different from each other with respect to the study design, the target populations, and the frameworks. Both are aimed at assessing the performance of education systems, and both achievement frameworks are organized around two dimensions: a content dimension specifying the content to be assessed and a cognitive dimension specifying the thinking processes to be assessed. But, while the TIMSS study is designed to reflect three levels of countries' curricula—the intended, the implemented and the achieved—the PISA study is designed to reflect science in everyday life or out-of-school situations. The PISA science achievement construct is based on an idea of various natural science systems in the real world (physical systems, life systems, earth and space systems). The science construct in TIMSS is instead based on the traditional disciplinary subdomains within the science school subject. In eighth grade these are biology, physics, chemistry, and earth science, and in fourth grade these are life science, physical science, and earth science.

The main difference appears however to lie in the selection of and construction of test items. In PISA, the science test tasks are extensive, text-rich, and context-specific. In TIMSS grade 8, the task stimuli are mostly short and contain purely scientific applications. Furthermore, in TIMSS more effort is made to cover and balance the subdomains within the science subject than is PISA. The science achievement scores from the two studies have however, despite these differences, been found to correlate highly at the country level (e.g., Klieme, 2016).

Gender Gaps in Science in Fourth-Grade Students

In the earliest IEA study of science in 1970, boys in most countries and both age-levels achieved a higher mean on the science scale than girls, and in the second in 1984, the gap was either smaller or nonexistent in many countries of the 1970 study (Keeves, 1992).

In TIMSS 2015, 25 out of 49 participating countries in grade 4 reported no gender gap in science achievement. In the other half, gender differences went in both directions, but the advantage was smaller in countries where boys had higher means, and vice versa. In grade 4, the mean score for boys was higher in 11 countries, with an average difference of 0.08 SD. In 11 other countries, the mean score of girls was higher, with an average difference of 0.24 SD (Martin et al., 2016). The gender gap among fourth-graders in TIMSS 2015 is depicted in Figure 4. In TIMSS 1995, when 25 country participated, all but New Zealand, Latvia, and Thailand reported higher science means for boys. Ten of those were statistically significant with a gap size that varied between 0.13 and 0.26 SD (Mullis et al., 2000).

The IEA has provided the opportunities to monitor within-country trends in gender gaps over time in grade 4 science achievement, since the TIMSS assessment has been repeated with comparable achievement measures since 1995. Seventeen countries participated in the cycles between 1995 and 2015 (Mullis, Martin, & Loveless, 2016). Ten of these countries showed a small but significant average advantage for boys of about 0.09 SD. By 2015, the number of countries with significant advantages of boys decreased to seven, and also the size of theses difference had decreased from 0.09 to 0.03 SD. Out of the 42 countries that participated at least in two cycles, nine never observed any gender differences, in seven the difference diminished, and in another five the difference has decreased. A small but stable mean advantage for boys can be observed in six countries (Chinese Taipei, Republic of Korea, Hungary, Slovak Republic, Italy, and Portugal). In another seven countries, a stable, emerged, or increased difference to the girls' advantage can be observed (Bahrain, Finland, Islamic Republic of Iran, Kazakhstan, Morocco, Oman, and Saudi Arabia).

It is clear that the distributions of science gaps between boys and girls in grade 4 across countries have changed quite a bit during this 20-year time period. Not only has the advantage of boys often diminished or decreased, but some countries also report a relatively substantial mean advantage of girls. Sixteen out of 42 countries with trend data have, by 2015, reported gender equality with respect to average achievement in science grade 4.

In summary, the main observation from TIMSS 2015 and previous cycles is that the gender gap in science has either decreased, diminished or in some cases turned towards a female advantage (Mullis, Martin, & Loveless, 2016). The same pattern was also observed in the TIMSS 2019 report (Mullis et al., 2020).

Gender Gaps in Science in Eighth-Grade Students

In TIMSS 2015, 20 out of 39 countries reported no gender gaps in science in eighth-grade students. In the other half, and similar to grade 4, the gender gaps were found to go in both directions. The magnitude of the difference was smaller in countries where boys had higher means, and vice versa. In 14 countries, a female advantage was observed with an average of 0.28 SD. A smaller male advantage was reported in five countries (on average 0.11 SD). Notable is that the mean advantage for girls was observed mainly in Middle Eastern countries such as Saudi Arabia, Bahrain, Kuwait, Oman, Jordan, United Arab Emirates, and Qatar, while male advantages were observed in Italy, Hungary, and Hong Kong. The countries that participated in both the grade 4 and the grade 8 assessment offer a possibility to observe differences between grade 4 and grade 8. In TIMSS, the pattern appears as follows: While the boys' advantage seems to be stable between grade 4 to grade 8 in Hungary and Italy, it seems to disappear in Chinese Taipei and the Republic of Korea. The girls' advantage found in the Middle Eastern countries in the grade 4 study appears to remain in all countries but Kazakhstan and the Islamic Republic of Iran.

Since the TIMSS assessment has been repeated every four years with comparable measures since 1995, one can monitor within-country trends in gender gaps over time in grade 4 science achievement. In the TIMSS 2015 report on science achievement (Martin et al., 2016), the changes in gender gaps across 20 years are presented. Out of 56 countries that participated in the eighth-grade TIMSS science assessment at least twice during this time span, only three countries (Kazakhstan, the Republic of Korea, and South Africa) never reported any statistically significant gender gap in science. Instead, a previously reported male advantage has disappeared in 16 countries, and in another three countries, the gap decreased from medium to small. A small but stable mean advantage for boys can be observed in Hungary and Italy, and for girls in Malaysia. A small but growing mean difference to the girls' advantage was observed in Lebanon, Malta, and Morocco. Lastly and more unexpectedly, a relatively large and increasing mean gap to the girls' advantage can be seen in nine countries: Bahrain, Jordan, Kuwait, Oman, Qatar, Saudi Arabia, Thailand, Turkey, and the United Arab Emirate (Martin et al., 2016).

The changes in the landscape of grade 8 gender gaps in science achievement since 1995 show in summary a disappearing male advantage, and in many countries, primarily in the Middle East, a rather large and growing female advantage.

Gender Gaps in Science in 15-Year-Old Students

In the first PISA assessment in 2000, gender gaps in science achievement were found to be small in most countries and sometimes in favor of girls and sometimes in favor of boys. In 24 countries, the gap was not statistically significant. In the primary report, a comparison is made to the TIMSS study from 1995 where the science gender gap was often in favor of males. The difference between the studies may, it is said, partly be explained by the content and form of the two studies with more emphasis on life science, on process and application, and more open-ended questions in PISA (OECD, 2001).

In PISA 2006, science was in the main focus for the first time and the data from that year were more closely analyzed from a gender perspective in a special report (OECD, 2009). Across countries, the overall gender difference was small with 0.02 SD in favor of boys. Six countries showed significant differences in favor of boys and two in favor of girls. For the content-demanding identifying scientific issues subscale, girls tended to perform better whereas boys scored higher at explaining phenomena scientifically. Countries were, in this report, categorized in three different types in terms of gender patterns found at the country and school level. The first type showed virtually no differences overall, or only small ones at the school level. In the second category, there were no significant overall gender gaps, but large within-school differences. The third pattern showed gender gaps both overall and within schools. Differences between these categories of countries can likely be, at least partly, explained by differences in the school systems (e.g. early versus late tracking), but there may also be other explanations. Several factors that could be expected to explain the differences, such as the large differences in choices made for tertiary education (e.g., Stoet & Geary, 2018, 2020), with for instance very few females choosing computer science and very few males choosing nursing does not mirror the small achievement differences seen at secondary school level. School types are also brought into the picture and results from earlier studies showing single-sex schooling to benefit girls is not confirmed in this study (OECD, 2009).

In PISA 2015, where science was in focus for the second time, boys scored better in 24 and girls in 22 out of overall 72 participating countries (OECD, 2016). More boys than girls were among the high-performers (at or above performance level 5), but, at the same time, there are fewer girls at the lower end of the distribution. An analysis of boys' and girls' relative strengths showed small differences with boys performing better on scientific explanations and girls on scientific design. These differences were larger at the top level of the distribution.

The PISA assessment from 2018 showed similar results with very small overall differences (OECD, 2019b). At the country level, no differences were observed in about half of the participating countries. As displayed in Figure 4, girls performed significantly better in 34 and boys in six countries. Most of the countries in which boys performed better were placed in Latin America. Countries in which girls performed better were heterogeneous in their geographical location and income levels.

Overall, science achievement, as measured in the PISA assessments, shows comparably small gender differences. This finding is similar to the TIMSS studies in grades 4 and 8. Overall and across the different studies, some substantial changes have been observed since 1971. Overall, the findings of the different studies indicate that the rather consistent male advantage in the early studies changed to a more or less closed or even reversed gender gap. Most of this change apparently took place before 1995. It seems that science could no longer be considered a "boy subject" in the school context.

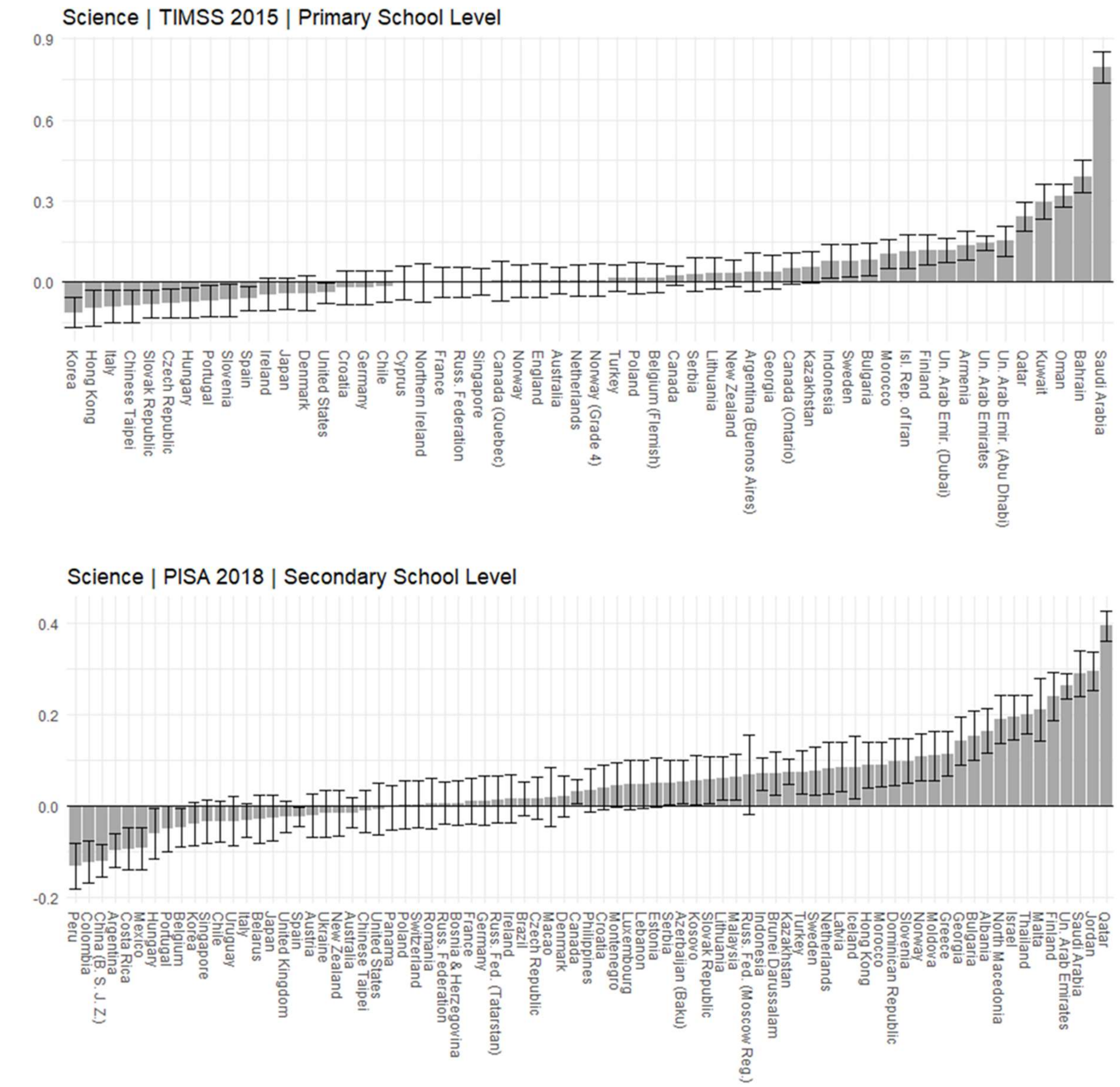

Figure 4. Gender gaps in science in units of the international standard deviation (SD; y-axis) in all participating countries and benchmarking participants in two recent ILSAs. Values

above zero indicate higher scores of girls compared to boys. Larger confidence intervals indicate smaller sample sizes.

Gender Gaps in Civic and Citizenship

Civic and citizenship knowledge contains facts about politics, social structures, and other societal issues, but the readiness to take an active part in civic activities is also considered important. The premises for the ICCS assessment was described as follows: "[…] preparing students for citizenship roles involves helping them develop relevant knowledge and understanding and form positive attitudes toward being a citizen and participating in activities related to civic and citizenship education" (Schulz et al., 2010a, p. 15). While, for instance, mathematics knowledge is mostly value-free, knowledge and understanding concerning civics and citizenship contains values and an understanding of how to act in society. Since political situations and values vary over time and space, as does the organization of societies, the content of civic knowledge also varies (cf. Schulz et al., 2010a, p. 22). What knowledge implies in this field is partly political and comparisons between countries may be controversial. However, comparative assessments are conducted, and civic and citizenship knowledge is important at the social and individual level. Knowledge and attitudes concerning gender equality are part of civics and citizenship education and may affect how boys and girls perceive their educational opportunities.

As a background to the more recent studies, it can be noted that an early assessment conducted by the IEA in 1971 showed that 14-year-old boys were more knowledgeable about societal issues and phenomena than females of the same age (cf. Schulz et al., 2010b). The results of the CIVED study from 1999 (Torney-Purta et al., 2001) measuring political content knowledge and skills in interpreting political communication showed no gender differences among lower secondary students in all but one of the participating countries. The exception was Slovenia where girls had significantly better results.

In the ICCS assessment from 2009 (Schulz et al., 2010a), girls showed statistically significantly higher average scores than males both at the international average and in nearly all participating countries. There was no association between the magnitude of the gap and overall performance levels or the geographical locations of countries. In the next assessment in 2016 (Schulz et al., 2018), the gender pattern in all but three of the 24 participating countries or regions showed higher scores for girls than for boys (see Figure 5): "Across all countries, the difference in average civic knowledge scale scores between female and male students was equivalent to roughly one-third of a level on the ICCS scale" (Schulz et al., 2018, p. 41).

In the ICCS study in 2009, one of the measured attitudes concerned gender equality. Even if most student were supportive of gender equality and equality in other dimensions, females were statistically significantly more so in all countries (Schulz et al., 2010a). This pattern was repeated in the 2016 assessment in questions about students' engagement in civic issues (Schulz et al., 2018). Other questions measured values and attitudes concerning, for instance, democracy, freedom of the press, and the state displayed no gender differences in any of the two assessments. But in 2009, female students in all countries were more likely to say that they were willing to do voluntary work to help others. The perceived ability to influence school processes showed statistically significant differences in less than half of the countries in 2009, most often to the advantage of males. In 2016, female students were more willing to participate in school issues.

It can be concluded that over the assessments of young teenagers' (14 to 15 years old) civic knowledge conducted by the IEA between 1971 and 2016, that is, 45 years, the gender gap has changed from males being more knowledgeable in the beginning of the period to girls being more knowledgeable in the most recent study.

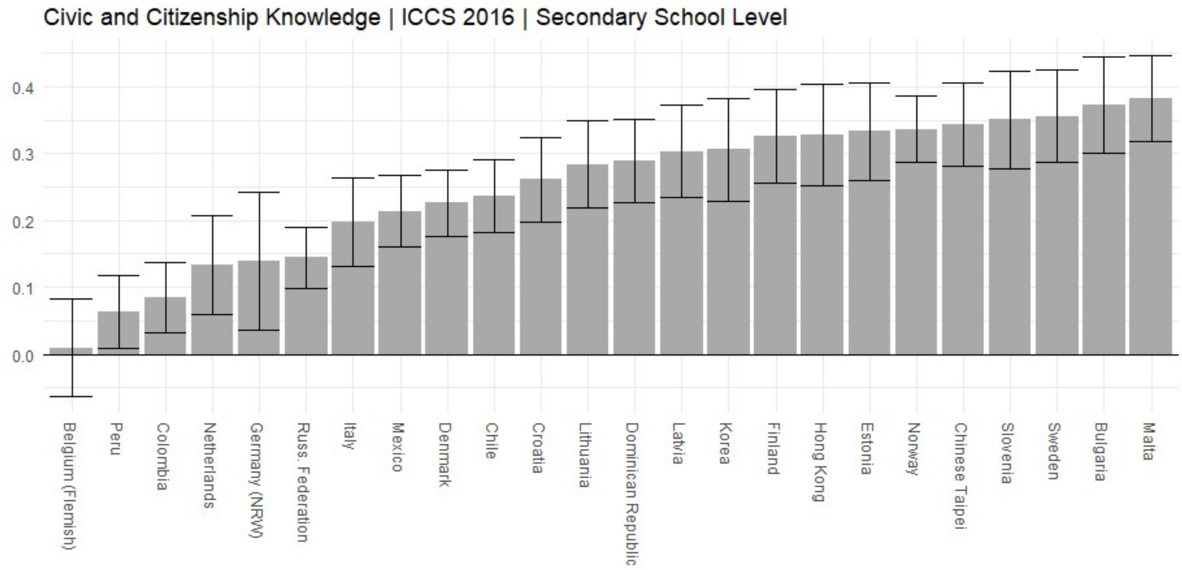

Figure 5. Gender gaps in civic and citizenship knowledge in units of the international standard deviation (SD; y-axis) in all participating countries and benchmarking participants in two recent ILSAs. Values above zero indicate higher scores of girls compared to boys. Larger confidence intervals indicate smaller sample sizes.

Gender Gaps in Computer and Information Literacy

The IEA launched the ICILS assessment in 2013 with the aim to assess the ways in which young people are developing computer and information literacy (CIL). The CIL construct combines information literacy, critical thinking, technical skills, and communication skills applied across a range of contexts and for a range of purposes. The ICILS 2013 study (Fraillon et al., 2014) was designed to describe international differences among grade 8 students and included 21 educational systems (18 high income and 3 middle income systems). Girls tended to outperform boys in all participating countries. Across countries, the average difference between females and males was 0.18 SD to the female advantage. The statistically significant gap within countries ranged from 0.12 SD in the Czech Republic to 0.38 SD in Korea.

In a report especially devoted to this unexpected gender pattern in the ICILS 2013 data (Gebhardt et al., 2019), potential student and teacher explanatory factors were explored in those 14 education systems which reported sufficient sampling requirements. A higher female average was reported for the 12 included high income countries, whilst in the upper-middle income countries Thailand and Turkey, the differences were negligible and the general achievement levels in CIL were very low. In this report, hypotheses of differences in student CIL self-efficacy, attitudes toward, interest in, and use of information and computer technology were investigated. No gender differences in self-efficacy with respect to CIL were found, but for some specific tasks, males rated their ability significantly higher than did females. Neither could the study reveal any consistent gender differences in information and computer technology interest or enjoyment. Some differences were found with respect to the use of productivity applications, where females reported more social, and males more recreational activities. None of these gender differences in interest and use could explain any part of the gender gap found in CIL achievement. The in-depth analysis of teachers' experience, disposition towards, or extent of pedagogical use of information and computer technology revealed no consistent differences between males and females across countries, either. The inability to link teacher characteristics and/or information and computer technology practices to student CIL achievement may however be due to limitations in the study design. The only variable with some explanatory value, was that females on average used a little more time to solve the test tasks. It was suggested that this might reflect more careful and thoughtful responses. Differential item analysis indicated that female students achieved relatively better in comparison to male students on tasks that involved communication, design, and creativity, whilst the opposite was found for more technically oriented tasks. The authors suggested that computational thinking could be a construct that would produce a different gender pattern.

In ICILS 2018 (Fraillon et al., 2020), student CIL was assessed again, with a design that included computational thinking (CT). This was defined as the type of thinking used when programming on a computer or digital device. The CT scale was based on the performance on two different computer programming tasks. In ICILS 2018, 12 countries and two benchmarking units (Moscow in the Russian Federation and North Rhine-Westphalia in Germany) participated. All but one (Kazakhstan) were high income countries.

Female gender tended to be positively related to CIL scores, but negatively related to CT scores (see Figure 6). In ICILS 2018, the average CIL scores of girls were statistically significantly higher than those of male students in all countries and benchmarking participants except Chile, Uruguay, and North Rhine-Westphalia (Germany). The average difference amounted to 0.18 SD. The female advantage on CIL ranged from 0.06 SD in Moscow (Russian Federation) to 0.39 SD in Korea. Four countries also participated in ICILS 2013, and compared with the 2018 study, the female advantage on CIL remained unchanged in Denmark, Germany and Korea, whilst the female advantage was no longer statistically significant in Chile.

Across the nine systems that participated in the optional CT assessment, the average scale score of male students was higher than that of females, but at the country level not consistently so. The male advantage was on average small and only statistically significantly in Portugal and the benchmarking system North Rhine-Westphalia (Germany). In Finland, the pattern was opposite, females performed significantly better than males (see Figure 6). In France, Germany, Korea, and Luxembourg, the average scores of male students were higher than those of female students but not statistically significant, and in Denmark, no mean gender difference was found at all.

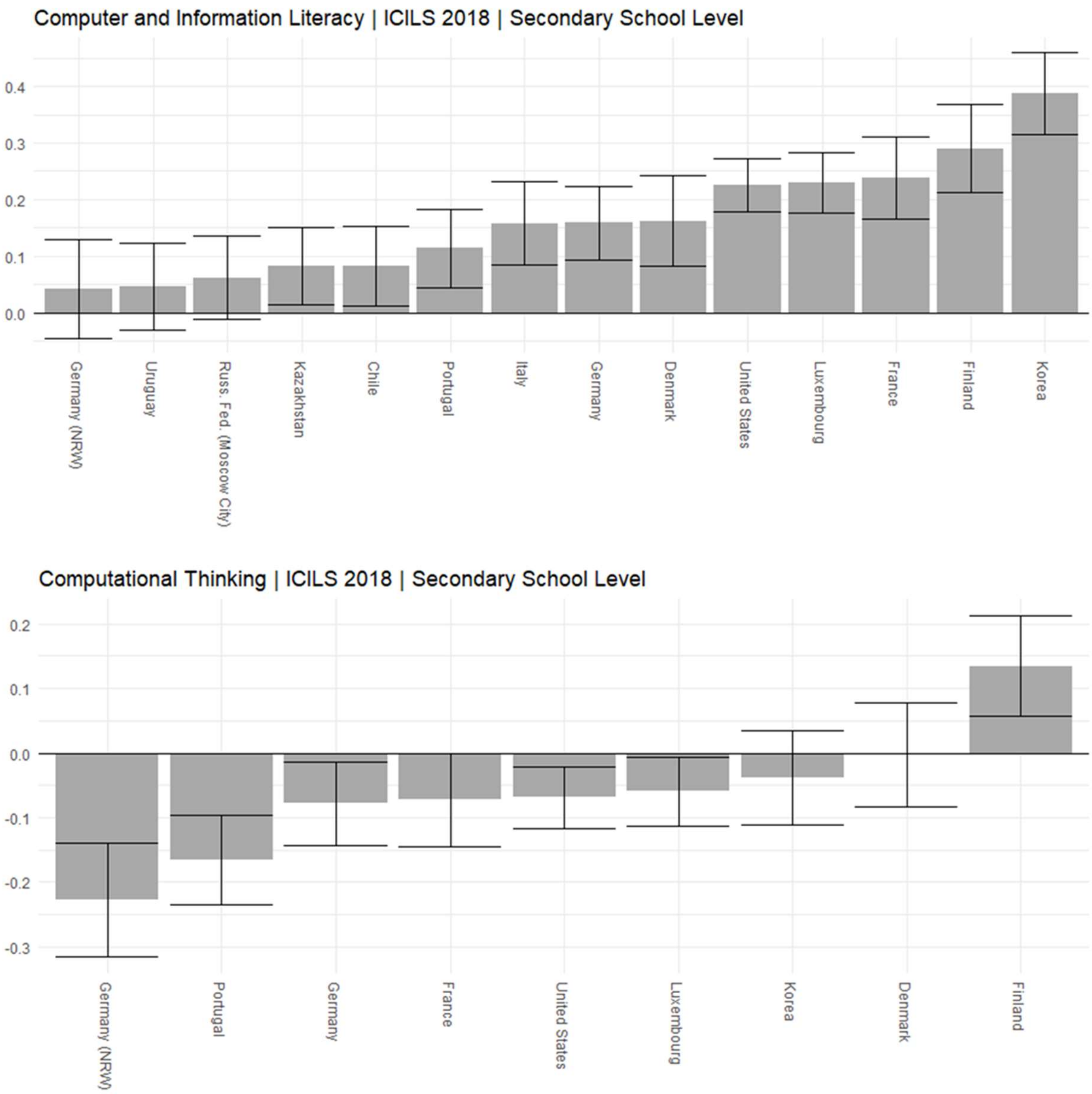

Figure 6. Gender gaps in computer and information literacy and computational thinking in units of the international standard deviation (SD; y-axis) in all participating countries and benchmarking participants in two recent ILSAs. Values above zero indicate higher scores of girls compared to boys. Larger confidence intervals indicate smaller sample sizes.

Secondary Analyses on Gender Differences in ILSAs

In the first part of the chapter above, the gender gap findings of the ILSAs, as summarized in the primary reports of the IEA and OECD, were overviewed. In the following second part of the chapter, the results of the systematic literature review on secondary analyses that used data from the ILSAs are presented. Overall, we included 42 studies that investigated associations between gender gaps in achievement and variables on the student, teacher, school, and/or system level that may or may not be associated with these gaps (see Table 1). In the following, we briefly outline key research questions and rationales and summarize the most central findings for the different achievement domains, separately. Our search revealed 23 secondary analyses on gender gaps in reading that used PIRLS or PISA data, 31 in mathematics that used TIMSS or PISA data, 16 in science that used TIMSS or PISA data, one on civic and citizenship achievement that used ICCS data, and none in computer and information literacy. In the domain-related subsections, we grouped the studies by the types of research questions.

Table 1

Overview of the 42 included studies

#	Study	Achievement domain					Level of gender gap research question			
		Reading	Math.	Science	Civic & citizen.	Comp. & inf. lit.	Student lev.	Teacher & school lev.	Country lev.	Variation in scores
1	Ababneh and Abdel Samad (2018)			X					X	
2	Ayalon and Livneh (2013)		X						X	
3	Baye and Monseur (2016)	X	X	X						X
4	Borgonovi (2016)	X					X			
5	Chamberlain (2003)			X					X	
6	Cheema and Galluzzo (2013)		X				X			
7	Cheung et al. (2013)	X					X			
8	Chiu (2010)			X					X	
9	De San Roman, Ainara Gonzalez and La Rica (2016)	X	X				X		X	
10	Dronkers and Kornder (2015)	X	X						X	
11	Else-Quest et al. (2010)		X				X		X	
12	Ghasemi et al. (2019)		X						X	X
13	Gray et al. (2019)	X	X	X						X
14	Guiso et al. (2008)	X	X						X	
15	Hamamura (2012)		X						X	
16	Hermann and Kopasz (2019)	X	X	X					X	
17	Hosenfeld et al. (1999)		X					X		
18	Ireson (2017)		X	X					X	
19	Kim and Law (2012)		X					X		
20	Machin and Pekkarinen (2008)	X	X							X
21	Mak et al. (2017)	X						X		
22	Marks (2008)	X	X					X	X	
23	Meelissen and Luyten (2008)		X				X			
24	Munir and Winter-Ebmer (2018)	X	X						X	

		Achievement domain					Level of gender gap research question			
#	Study	Reading	Math.	Science	Civic & citizen.	Comp. & inf. lit.	Student lev.	Teacher & school lev.	Country lev.	Variation in scores
25	Neugebauer et al. (2011)	X	X	X				X		
26	Nosek et al. (2009)		X	X					X	
27	Penner (2008)		X						X	X
28	Pereira et al. (2015)				X		X			
29	Reilly (2012)	X	X	X					X	
30	Reilly et al. (2019)		X	X					X	X
31	Rodríguez-Planas and Nollenberger (2018)	X	X	X					X	
32	Schwabe et al. (2015)	X					X			
33	Shera (2014)	X						X		
34	Smith et al. (2017)	X	X	X			X			
35	Stoet and Geary (2013)	X	X						X	
36	Tao and Michalopoulos (2018)		X						X	
37	Tsai et al. (2018)	X	X	X				X		
38	van Hek et al. (2018)	X						X		
39	van Langen et al. (2006)	X	X	X					X	
40	Wiseman (2008)		X					X		
41	Yamamura (2019)	X	X	X					X	
42	Zhou et al. (2017)		X							X

Reading Achievement

Explanatory Variables at the Student Level

A first group of reading-related primary studies that our literature review identified investigated questions on the student level, that is, if reading gender gaps relate to differential characteristics or behaviors of boys and girls. One study tested if the generally higher intrinsic reading motivation of girls in comparison to boys would relate to the girls' particular strength in constructed response items (i.e., items in which students respond with free text) (Schwabe et al., 2015). The authors found this assumed association in PISA but not PIRLS data from Germany. Girls generally reported a higher intrinsic reading motivation than boys, which was associated to higher reading achievement independent of the item format (ibid.).

Two studies addressed gender gaps in reading digital texts: Borgonovi (2016) used PISA data from 26 countries and generally found gender gaps in favor of girls in both reading printed and digital texts. The female advantage in reading digital texts was however not as pronounced as the one in printed texts. This study investigated if this smaller disadvantage of boys in digital texts could relate to differential video gaming behavior of boys and girls and found some empirical support for their assumptions. A second study by Cheung et al. (2013) also investigated reading digital versus printed texts and used PISA data from Hong Kong and Korea. Girls scored higher on the digital reading test than boys in both countries. The two countries varied regarding the reported gender differences in information and computer technology use at home, which was associated to differences in digital reading achievement. However, the direction and sizes of these mediation effects varied across countries and types of computer-related activities.

In two studies, the authors examined whether the relationship between family background and reading achievement differed between boys and girls. De San Roman, Ainara Gonzalez and La Rica (2016) used 64 country samples from PISA and found that having a mother participating in the labor force was positively associated with reading achievement for both boys and girls. This positive relationship was however found to be slightly more pronounced for girls in comparison to boys within countries. This might indicate that especially girls profit from a family with maternal labor force participation. Another study that used PIRLS 2011 data from Germany and investigated if the familial economic and cultural background would relate differently to reading achievement for boys and girls did however not find such differential associations (Smith et al., 2017). In other words, the socio-economic family background seemed to be equally important for the reading acquisition of boys and girls.

Explanatory Variables at the Teacher and School Level

A second group of studies focused on relationships between reading gender gaps and variables at teacher and school levels. One study addressed the long-standing assumption that boys could learn better from male teachers and girls could learn better from female teachers. Since most primary school teachers are usually female, this argumentation is often considered to explain girls' reading achievement advantages. The study by Neugebauer et al. (2011) used data from PIRLS and a national extension in Germany and investigated the relationship between having teachers with the same versus not the same gender and student reading achievement.

They did not find achievement differences between students who were taught by teachers with the same versus with the opposite gender, after controlling numerous further variables.

Five studies investigated if the female reading advantage would relate to a segregation of boys and girls into different schools, that is, girls more frequently attending schools with characteristics that foster learning. Mak et al. (2017) used data from Macao in PISA and found that the girls' advantage in reading was partly explained by the fact that they more often attended schools that were composed of students who enjoyed to read, and by having better reading-related attitudes and behaviors themselves. Similarly, Marks (2008) investigated 32 country samples from PISA and found that the segregation of boys and girls into different schools accounted for some of the reading gender gap, while this varied between countries. Shera (2014) used PISA data from Albania to investigate associations between the reading gender gap and school variables and found the gender gap to be larger at public in comparison to private schools, when controlling for the schools' reading levels and location. Similarly, van Hek et al. (2018) studied the association between reading gender gaps and school characteristics in 33 countries in PISA. They found that schools with a high share of female students showed higher reading scores, especially of boys. Due to the cross-sectional design of the study, one cannot conclude that attending a school with a majority of girls would be especially reading promoting for boys. It could for instance also be assumed that students with high reading abilities are selected into the same schools, so that more girls and fewer but high performing boys become classmates. Schools with many students with highly educated parents and with many highly educated teachers showed equally higher reading scores for both genders. In contrast, Tsai et al. (2018) rather found that family and school characteristics affected achievement equally for boys and girls, that is, did not account for gender gaps in reading achievement. They used PISA data from Japan, South Korea, Taiwan, USA, Germany, and the Czech Republic.

Explanatory Variables at the Country Level

A third group of studies investigated associations between country-level variables and reading gender gaps. Seven studies investigated if the equality of men and women in society would also reflect in differential reading gender gaps in school children. These studies all used PISA data from different countries and cycles, as well as a plethora of different control variables and measures of equality of men and women on the country level. These measures for example depicted equalities in the labor force participation, income, and political empowerment. While most studies found the reading advantage of girls to be higher in more gender equal societies (De San Roman, Ainara Gonzalez & La Rica, 2016; Guiso et al., 2008; Marks, 2008; Reilly, 2012; van Langen et al., 2006), two studies found mixed or insignificant results (Munir & Winter-Ebmer, 2018; Stoet & Geary, 2013). This kind of investigations of relationships between societal gender equality and gender gaps in student achievement therefore address the assumption that societal values matter for student learning, for example through schools, teachers, parents, etc.

Three studies also examined societal measures of gender equality and their associations to reading gender gaps but focused on immigrant students. These argued that the effects of the societies of origin and the effects of the societies of residence could be meaningfully disentangled for immigrant students. They therefore used subsamples of immigrant students in

PISA data from different countries and addressed if the socialization of parents in the countries of origin matters over and beyond the society in which the children go to school. Dronkers and Kornder (2015) used PISA data and investigated the associations between the societal gender equality levels of the countries of origin and reading achievement for immigrant boys and girls. They found that immigrant girls showed higher reading scores than comparable immigrant boys, if their countries of origin had more gender equal societies. The gender equality level in their countries of residence, on the contrary, related to reading achievement equally for boys and girls. Rodríguez-Planas and Nollenberger (2018) focused on nine countries of residence in PISA and immigrant students from overall 35 countries of origin. They found that immigrant students from more gender equal countries of origin generally showed higher reading scores in their countries of residence and that this effect was more pronounced for girls than for boys. Yamamura (2019) also investigated immigrant students in PISA, using samples from 63 countries. This study did not find a significant association between the gender wage gap in the country of residence and the reading gender gap in immigrant students.

Two further studies investigated if the differentiation in school systems, that is, the allocation of students to different school tracks and types relates to countries' gender gaps in reading achievement. A study by van Langen et al. (2006) analyzed data of 42 countries in PISA and found the girls' advantage in reading to be more pronounced in education systems with a low level of differentiation and in rural in comparison to urban areas. Hermann and Kopasz (2019) used data of numerous countries in PISA and PIRLS and found that the reading gender gap was more pronounced in fifteen-year-olds than in fourth-graders in all countries, except Great Britain. The increase of the reading gender gap between primary and secondary school students was more pronounced in countries with early school tracking policies than in late-tracking countries.

Studies on Gender Gaps in the Variation of Reading Scores

Three studies did not focus on gender gaps in mean achievement levels but on the differences in the distribution of reading scores between boys and girls. Using PIRLS and PISA data from different numbers of countries, all three studies showed that the variability of reading scores was on average larger in boys than in girls (Baye & Monseur, 2016; Gray et al., 2019; Machin & Pekkarinen, 2008). Especially in the lowest performing part of the distribution, boys were highly overrepresented (Baye & Monseur, 2016; Gray et al., 2019).

Mathematics Achievement

Explanatory Variables at the Student Level

A first group of mathematics-related studies that the systematic literature review identified investigated associations between mathematics achievement gaps and student or family characteristics. One group of these studies emphasized the interplay between mathematics achievement, gender, and mathematics-related attitudes. Different studies found boys to report higher mathematics self-confidence scores than girls (Cheema & Galluzzo, 2013; Else-Quest et al., 2010; Meelissen & Luyten, 2008). When comparing students with the same mathematics self-confidence levels, the achievement score advantage of boys diminished in a study that focused on US-American students (Cheema & Galluzzo, 2013), and even turned into an advantage of girls in a study that focused on Dutch students (Meelissen & Luyten, 2008).

Two studies focused on the family background. Just as for reading above, De San Roman, Ainara Gonzalez and La Rica (2016) found that having a mother participating in the labor force was positively associated with mathematics achievement for both boys and girls. This relationship was slightly more pronounced for girls in comparison to boys which could indicate that especially girls profit from mothers who work outside the home. Smith, Wendt, and Kasper (2017) focused on the economic and cultural background of the students' families and found, just as for reading, no support for the assumption that the positive relationship between these variables and mathematics achievement would differ between boys and girls.

Explanatory Variables at the Teacher and School Level

A number of studies focused on variables on the teacher or school level and their role for gender gaps in mathematics. One group of these studies assumed that boys and girls are, through various processes, selected into different schools and that these differences account for parts of the gender gaps in achievement. While Tsai et al. (2018) found that school characteristics related equally to mathematics achievement for boys and girls, they also found that selection mechanisms could account for a share of the mathematics gap in some countries. Hosenfeld, Köller, and Baumert (1999) used a longitudinal extension of a TIMSS assessment in Germany to investigate mathematics gender gaps in different school types. They found that girls more often attended higher school tracks than boys. There was no overall mathematics gap between boys and girls across all schools, but compared with girls from the same school types, boys attained higher scores. One other study that used PISA data from 32 countries did not find that the gender segregation of boys and girls in different schools would significantly explain the gender gap in mathematics, unlike the one for reading (Marks, 2008). Similarly, Wiseman (2008) looked at the countries' proportions of gender-segregated mathematics classrooms, and found no clear linear association with the countries' mathematics gender gaps in TIMSS. Also, Kim and Law (2012) found mixed results for the association between gender segregated and joint schooling and mathematics achievement in Korea and Hong Kong.

Another study used data from a national extension of a PIRLS assessment in Germany to investigate the relationship between mathematics achievement and their teachers' gender. They did not find achievement differences between students with the same versus opposite gender in mathematics, just as in reading (Neugebauer et al., 2011).

Explanatory Variables at the Country Level

Third, a number of studies investigated the association between societal gender disparities and gender gaps in the mathematics achievement scores of school children. Just as in the case of reading achievement above, these studies argue that gender-related societal values might matter for boys' and girls' learning processes, for example through schools, teachers, parents, etc. These studies investigated the associations between countries' mean gender gaps in mathematics and indicators of gender-related societal norms and processes, such as the degree of gender equality on the labor market, in politics, and health. Four of these studies found that the relative performance of girls compared to boys was higher in more gender-equal countries (De San Roman, Ainara Gonzalez & La Rica, 2016; Guiso et al., 2008; Marks, 2008; Reilly, 2012). In other words, these studies found boys and girls to perform about equally well or girls to perform better than boys in more gender-equal countries, while boys outperformed girls in less gender-equal countries. In contrast to the findings for reading, more studies found no significant or mixed relationships between measures of societal gender equality and mathematics gender gaps, however (Else-Quest et al., 2010; Ghasemi et al., 2019; Hamamura, 2012; Ireson, 2017; Munir & Winter-Ebmer, 2018; Penner, 2008; Reilly et al., 2019; Stoet & Geary, 2013; Tao & Michalopoulos, 2018; van Langen et al., 2006). Another study investigated the association between gender gaps and a more proximal measure of gender-related stereotypes. For 34 countries, TIMSS 2003 grade eight data was combined with information about the degree of stereotypical beliefs about science being a more male domain and liberal arts being a more female domain (Nosek et al., 2009). The countries' mean stereotypical beliefs were assessed in a large international implicit association test study. This study found that more pronounced gender-related stereotypes in the society correlated with larger mathematics achievement advantages of boys as compared to girls. Together, these studies draw a mixed picture about the role of gender-related societal norms and processes for mathematics achievement gaps. The choice of indicators for gender-related societal characteristics seems to matter for understanding gender gaps in mathematics.

Another group of studies that investigated societal country-level explanatory variables focused on the subgroup of immigrant students. Just as for reading above, these studies compared mathematics gender gaps between immigrant students from different countries of origin. Rodríguez-Planas and Nollenberger (2018) as well as Yamamura (2019) found that in groups of students who immigrated from more gender-equal societies, girls scored equally high or higher than boys in mathematics, while in groups from less gender-equal societies, boys outperformed girls in mathematics. While these studies support that gender-related values are important for gender gaps, such effects were not significant in a similar study by Dronkers and Kornder (2015).

Other studies investigated associations between mathematics gender gaps and more structural and political characteristics of educational systems. One study investigated the role of early as compared to late tracking policies and found, in contrast to the above-mentioned effects on reading gender gaps, no significant effects for mathematics (Hermann & Kopasz, 2019). Van Langen, Bosker, and Dekkers (2006) used a broader definition of educational differentiation that for instance included tracking and between-school differences in the socio-economic composition of the student body. Based on PISA data of 42 countries, they found the relative performance of girls compared to boys to be higher in rural than in urban areas and in less differentiated school systems. In other words, in urban areas and highly differentiated school systems, they found larger mathematics gaps in favour of boys. Ayalon and Livneh (2013) did not focus on the role of educational differentiation but educational standardization for mathematics gender gaps. Using TIMSS data from 32 countries, they found lower mathematics gender gaps in countries with central exams as compared to countries without central exams, and in countries in which the teachers reported more homogeneous instructional behaviour as compared to countries with heterogeneous teaching. Even though these findings do not allow conclusions about possible causes of gender gaps, they emphasize that educational gender gaps are embedded in a complex interplay of societal and political structures.

Studies on Gender Gaps in the Variation of Mathematics Scores

Just as in the case of reading achievement above, a set of studies investigated if boys and girls differ in the variability of test scores in mathematics. Most of these studies found a (slightly) larger variability of boys' scores as compared to girls' scores (Ghasemi et al., 2019; Gray et al., 2019; Reilly et al., 2019) or an overrepresentation of boys among high performers (Baye & Monseur, 2016; Machin & Pekkarinen, 2008; Zhou et al., 2017). An early study by Penner (2008) that used TIMSS data from students in the last year of secondary schooling from 1995 did not find substantive differences in the mathematics score distributions of boys and girls, though.

Association between Reading and Mathematics Gender Gaps

Four papers additionally studied the relationship between countries' gender gaps in reading and mathematics achievement. All four studies found the gender gaps in reading and mathematics achievement to correlate. The higher the advantage of girls in reading in a country, the better the girls' mathematics scores in relation to boys. This implies that countries with a high reading advantage of girls showed no significant mathematics gender gap or mathematics advantages of girls, while countries with small reading advantages of girls showed larger mathematics gender gaps in favor of boys (Guiso et al., 2008; Marks, 2008; Stoet & Geary, 2013; van Langen et al., 2006).

Science Achievement

Explanatory Variables at the Individual and Family Level

Family background, in terms of economic and cultural resources, was used in a study of Smith et al. (2017). They found that the family social background had the same consequences for science achievement for both boys and girls.

Explanatory Variables at the Teacher and School Level

A common hypothesis in the context of gender differences in school performance is that the teachers' gender would explain such differences to some extent (cf. Neugebauer et al., 2011). A same-sex teacher is often expected to have a more positive effect than a teacher of the other sex. Neugebauer, Helbig, and Landmann (2011) used data from IGLU-E, an expansion of PIRLS in Germany, to estimate whether there was a causal effect of having a same-sex teacher on student achievement. They compared different types of student outcomes (test scores and teachers' grades) in traditionally considered 'female' and 'male' subjects. Achievement differences between boys and girls could not be explained by their teachers' gender. Therefore, no support for the teacher-bias hypothesis was found. The authors concluded that "the popular call for more male teachers in primary schools is not the key to tackling the growing disadvantage of boys" (Neugebauer et al., 2011, p. 669).

Explanatory Variables at the Country Level

Another group of studies aimed to explain gender gaps with different measures of gender equality in a country. Economists hypothesize for instance that a diminishing gender gap in income would affect the achievement of girls. Yamamura (2019) tested this hypothesis. Reducing the sample to students with immigrant background, he singled out the effect of the size of the gender wage gap. He found associations between countries' mean science gaps and the wage gap between females and males. Hence, each 1 % decrease in the wage gap resulted in a 0.13 % increase in girls' science test scores. Yamamura (2019) claimed that the effect had both direct and indirect causes. An indirect effect was expected to result from frequency of truancy, since school attendance was thought to be related to the wage gap (ibid.). However, the claim of a causal relation between the size of the wage gap and school achievement may be overly optimistic. It seems unlikely that young girls have information about, and the ability to assess the wage gap and that they make attendance and school performance decisions by an economic rationality. However, gender values, beliefs, and structures are likely to influence both the wage gap and performance levels. This is indicated by the next study.

Rodríguez-Planas and Nollenberger (2018) also focused on second-generation immigrants because they share the institutions and culture of the country they live in but are also influenced by social gender norms from their parents' country of origin. This study also assessed reading and mathematics gender gaps (see above). Merged data from PISA 2003, 2006, 2009, and 2012 were used in connection with the World Economic Forum's gender gap index. One intention was to find out whether gender norms have consequences for science achievement. Gender norms in the parents' country of origin were found to affect girls' performance in science, that is, girls with parents from countries that were more gender equal performed better relative to boys. This evidence indicated that parents' expectations on girls' academic achievement have effects on cognitive skills but not on motivation and some other non-cognitive variables.

Even if it may be necessary to focus on one or a few variables at a time it is clear that gender differences have to be the result of complex relations between multiple factors. Chamberlain (2003) used TIMSS data from 1995 and 1999 for New Zeeland primarily to analyze if the gender differences were related to ethnicity. It was shown that gender differences in fourth-graders changed in different ways in two different ethnic groups (a male advantage disappeared in the Maori group, while a boy advantage in the Asian group increased). In grade 8, over the same period, the overall gender gap in New Zeeland decreased from one of the largest to one of the smallest (0.07 SD) internationally. This change was due to girls performing better and boys worse. It could be shown that this change was not present in all ethnic groups. One conclusion from the study is that subgroups in one country, as identified by ethnicity, can vary in results on gender difference in science performance.

Nosek et al. (2009) used TIMSS data from 2003 to analyze associations between countries' *implicit stereotypes* and gender gaps in science achievement. Implicit stereotypes were measured via the speed of, in this case, associations of science-related items with males or females. This data, collected independently from TIMSS in a virtual laboratory, included adult individuals from all over the world and showed that science was strongly associated with males in many countries. Gender differences in science performance were related to the average national implicit stereotype score for 34 countries. A correlation of $r = 0.60$ indicated that implicit stereotypes about science being a male domain in countries' adult populations were strongly related to gender differences in science achievement in favor of boys.

A number of studies correlated countries' science gender gaps with other measures of gender equality. Reilly (2012) compared gender differences in reading, mathematics, and science literacy in 65 countries in the 2009 PISA-assessment. More pronounced gender equality at the country level was related to smaller or no science gender gaps. The impact of the number of women in research positions was also analyzed. More women in science was associated with higher female performance on the PISA science test. Analyzing data from TIMSS 2011, Reilly et al. (2019) found that gender differences in science achievement and national levels of gender equality were associated and the performance of boys was more inconsistent than that of girls in most nations. Boys had more favorable attitudes towards science, and girls had lower self-efficacy beliefs. Another study analyzed PISA 2001 data and concluded that the national and between-school variation in gender differences was related to the structure of the school systems (van Langen et al., 2006). They also found a relation between a higher frequency of women in science, technology, engineering and mathematics fields in tertiary education and smaller gender gaps in science, driven by better performances of girls. In contrast to these findings, Ireson (2017) did not find any relationship between societal gender equality measures and gender gaps in science achievement using PISA data from 2012.

Chiu (2010) used data from PISA 2006 to test hypotheses on different relations between environmental awareness and science interest on the one hand and aspirations and achievement on the other. The study based on the idea that for boys, personal and social interest are integrated while for girls, they are contrasted, since girls would feel more responsible for the consequences of how science is used. They found some empirical support for these hypotheses. The conclusion was that country comparisons indicated that these kind of gender differences are induced by culture rather than biology.

Tsai et al. (2018) compared three types of school systems related to tracking, the use of centralized tests, cultural and expectations on achievement levels. The study compared three East Asian countries (low tracking), Germany and the Czech Republic (early tracking), and the USA (comprehensive system with within-school tracking). Using a multilevel multiple indicators multiple causes (MIMIC) model and PISA 2012 data, it was found that the gender gap in science differed depending on environmental or structural factors. There were no family or school effects on gender differences. When comparing with PISA 2001 data, a general conclusion, was that observed changes in gender differences depended on structural characteristics in the school systems, especially gender-segregating schooling.

Similar results were found by Hermann and Kopasz (2019) who explored whether between-country variation in science gender gaps were related to educational policies. They found relations between system-level factors and the size of gender differences. Early tracking was found to have a direct effect on the gender gap in favor of girls. This could be related to boys more often enrolling in vocational rather than academic tracks, where less time is spent on academic subjects. More student-oriented teaching practices (as perceived by students) were also associated with lower overall performance levels and a larger gender gap in favor of girls.

Ababneh and Abdel Samad (2018) focused on the gender gap in science in Jordan in PISA 2015. Jordan showed a very large gender gap in science in favor of girls across multiple PISA cycles. They concluded that system changes to close the gap have not been successful.

Studies on Gender Gaps in the Variation of Science Scores

One over many years discussed issue is the greater male variation hypothesis. It is often found in empirical research that more males score at the bottom and top ends of the achievement distributions. Gray and colleagues (2019) replicated and expanded an earlier study (Baye & Monseur, 2016) and used meta-analyses and meta-regressions to compare female and male variations over countries. They used data from PISA, PIRLS, and TIMSS between 1995 and 2015. For science literacy, they found that in 69 % of the countries (86), boys showed a larger variation in their science scores; in only two countries, girls showed a larger variation. They concluded that overall and in almost all countries, one can observe a greater male variability. The meta-regression investigated some possible explanations. Countries with higher performance levels and in which girls outperformed boys tended to show a larger variability among boys. Women's participation in society, as measured by the Global Gender Gap Index, tended to go along with an increased score variability of girls.

Civic and Citizenship Achievement

Civic and citizenship achievement is, as mentioned above, a subject with a somewhat different character. It contains important knowledge and skills to understand, analyze and make informed choices in the political sphere and regarding social science questions. However, the subject domain also contains attitudes and ideological stands that are not facts. Women's rights and gender equality are among the issues where there are profound differences between different cultures. Even if there is a strong and world-wide movement in the direction of gender equality, there are also stagnating or even reverse developments in others (e.g., UNESCO, 2019; World Economic Forum, 2019).

Pereira et al. (2015) used ICCS data from 2009 to analyze the gender differences in political knowledge more closely. This study started from earlier findings that boys tended to know more political facts than girls, but girls performed better when reasoning about political issues. The authors argued that different kinds of knowledge have to be distinguished to understand the gender dimension of civic knowledge. How political or civic and citizenship knowledge is measured is crucial to what gender differences are found. In this analysis, a distinction was made between factual and analytical knowledge. The authors also argue that by studying young people, gendered differences in adult experiences that affect knowledge can be avoided. The results indicate that boys were more knowledgeable about political structures whereas girls knew more about social and human rights. As expected, boys performed better on factual and girls on analytical knowledge domains (ibid.).

Computer and Information Literacy

Our search did not identify a secondary analysis on gender gaps in computer and information literacy that was conform to our inclusion criteria.

Discussion

Overall Results

The observed gender achievement gaps varied across domains and countries, but were rather stable over the investigated time period (i.e., 1995 to 2018). However, the first international assessments were already conducted in the 1970s, and displayed advantages for boys in all areas except of reading (mathematics, science, civic and citizenship). Therefore, the findings of the modern ILSAs since 1995 should be viewed against the backdrop of the more substantial changes that took place before 1995. When taking this long-term perspective of almost half a century of findings, is becomes evident that the female disadvantage has decreased, diminished, or even turned into a female advantage across the subjects. However, when reviewing these long-term international trends, it has to be considered that different countries participated in the different cycles and the assessments were not all directly comparable.

When summarizing the findings of the more comparable modern studies since 1995, a few robust patterns can be observed. In reading, girls outperform boys across almost all countries; especially in the low-achieving part of the distribution. In mathematics, the differences are small and sometimes in favor of girls and slightly more often in favor of boys in other countries. In science, there are no overall differences (i.e., girls perform better in some and boys in other countries) and in civic and citizenship, girls have an advantage in almost all countries. Girls showed better computer information literacy skills, but boys had higher scores on computational thinking in most countries (computational thinking was however only investigated once and in only nine countries).

Overall, gender gaps seem to be less pronounced in younger than in older students. Gender gaps also correlate across subjects. In countries with large advantages of girls over boys in reading, girls also show small advantages over boys in mathematics. In countries where girls have rather small reading advantages, boys typically show higher mathematics scores than girls. Generally, boys seem to show a larger variability of test scores in the different subjects, meaning that there are more boys both among those with very high and those with very low levels of performance.

The associations between gender gaps and other student characteristics (e.g., learning motivations, reading behavior) are rather stable as are associations between gender gaps and country characteristics like overall gender equality indicators. In such secondary analysis studies, results differed depending on the used ILSAs and country samples, as well as country-level gender equality measures (e.g., sub-facets of the World Economic Forum's Gender Gap Index). Some other associations between gender gaps and school characteristics were mixed, for instance single-sex schooling versus coeducation or different forms of school tracking did not relate to gender differences in a congruent way.

In the identified secondary analyses, many attempts were made to find explanations for specific results and to construct more comprehensive theories. There were attempts to construct reasonable models to understand the complicated gender patterns. Biological explanations, social structures and/or cultural factors were scrutinized by different authors. Most prominent were the gender similarities hypothesis, the gender stratification hypothesis, and the greater male variability hypothesis. Other highlighted issues were for example the stereotype threat (i.e. the assumption that the knowledge of stereotypes about being low-performing has a negative effect on actual test performances), the same-sex teacher bias (i.e. the assumption that girls learn better from female teachers and boys from male teachers), same-sex role models (i.e. the assumption that girls profit more from female and boys from male role models), or subject content that is culturally connected to males or females. Biological explanations lean on factors carried within the individual, but all other explanations present complex mixes of tendencies on structural, interactional, and individual levels (e.g., Halpern, 2012).

Methods of Analysis of the Secondary Studies

The secondary analyses of ILSA data on relationships between gender gaps in student achievement and variables on the student, teacher, school, or country level mostly used single-or multi-level regression analyses, structural equation modeling, or meta-analytical estimation tools in combination with meta-regressions. While such designs do not allow to draw conclusions about causal effects on gender gaps in achievement, they still illustrate interesting patterns of associations. Some of these studies did however use a language that did not acknowledge the limitations of cross-sectional approaches appropriately.

It seems worth to stress that we did not include studies that described gender gaps without relating them to external variables, as in Siddiq and Scherer (2019) or Meinck and Brese (2019) for instance, or that investigated associations between gender gaps and characteristics of the used tests (e.g., item format). We also excluded studies that regressed achievement on gender, while including numerous further covariates at once. If several covariates are added to a model in one step, changes in the regression coefficient of the gender variable cannot be attributed to single covariates. To allow such conclusions, covariates have to be added stepwise in separate models. We also excluded studies that investigated relationships between boys' and girls' achievement and predictors in parallel models, since these assume boys and girls to be separate populations and do not allow to draw conclusions about associations with gender gaps in achievement. In general, studies that explicitly specify interaction effects between gender and predictor covariates or that model the gender gaps as dependent variables directly seem to be more promising than these simple regression or parallel approaches. Also, it seems worth to stress that more research could profit from using multi-level modeling techniques.

Some studies used data from single ILSAs or combinations of them. Some used data from single, few, or many countries. Focusing on these ILSAs has several advantages, as the opportunity to compare countries, the large sample sizes and high-quality instruments. It does, however, also imply certain limitations. For example, the mentioned large-scale studies focus on students, i.e. children and adolescents that actually attend school. Therefore, the findings do not tell us anything about out-of-school boys or girls. Findings from household-based international assessments (see e.g., UNESCO, 2018) can complement those of the school-based assessments in many important regards, since gender equality in access to school and school attainment are key aspects of educational inequalities, apart from school achievement inequalities. Also, the participating countries can exclude part of the school populations, e.g., if they have special educational needs that prevent a full participation in the studies.

Last but not least, the presented ILSA studies and secondary analyses that base on them are restricted to the participating countries. Although the number of countries and world regions that participate in the international studies increases with time, it is important to remember that they do not reflect educational systems especially in lower income areas.

Implications of the Literature Review for Future Research

When finishing the literature review and summarizing the results, we were surprised by the scarceness of studies of interest. Especially ICCS, ICILS, and PIRLS appear underused to investigate gender gaps in student achievement, while TIMSS and PISA are used more often. ICCS and ICILS provide the unique opportunity to study gender gaps in civic and citizenship and computer and information literacy from a comparative perspective, two central outcomes of education in the 21st century. However, we only found one study that used ICCS (Pereira et al., 2015) and none that used ICILS. While reading gender gaps were more often analyzed in studies that used PISA data, PIRLS data appears underused. It seems valuable to investigate not only reading gender gaps in adolescents but also beginning readers in future research.

Furthermore, some research questions were investigated more often than others in the identified secondary analyses. One of the most interesting gender gap related finding of the presented ILSAs is the huge international variation in gender gaps in the different achievement domains. It therefore seems interesting that rather few studies investigated the associations of the countries' gender gaps with policy- or culture-related variables. We found four studies that hypothesized that gender gaps might vary between countries where boys and girls visit the same schools and classrooms and countries with gender-segregated education. They investigated whether tracking or gender-related differentiation correlates with gender gaps, assuming that school track or differentiation decisions might be gender biased (Ayalon & Livneh, 2013; Hermann & Kopasz, 2019; Tsai et al., 2018; van Langen et al., 2006). Future research should explore this possible explanation in more detail. The biggest group of identified studies correlated the countries' gender gaps with variables that indicate more general gender gaps in the societies (e.g., gender gaps in wages, political empowerment, health service access) (De San Roman, Ainara Gonzalez & La Rica, 2016; Else-Quest et al., 2010; Ghasemi et al., 2019; Guiso et al., 2008; Hamamura, 2012; Ireson, 2017; Marks, 2008; Munir & Winter-Ebmer, 2018; Penner, 2008; Reilly, 2012; Reilly et al., 2019; Stoet & Geary, 2013; Tao & Michalopoulos, 2018; van Langen et al., 2006). The choice of cultural variables, large-scale assessments, and countries varied however greatly between these studies, which limits the conclusions that can be drawn. Also, some of these studies controlled for instance for socio-economic differences between the countries, while others did not. Most of these studies modeled simple linear relationships between gender gaps in achievement and the more general society.

We believe that this field of country-level analyses on the variation of gender gaps between countries would benefit from more systematic future research. Naturally, country-level studies suffer from a low sample size of countries, which is why a systematic approach of using all available data from all participating countries and all assessment cycles seems to be a worthwhile avenue for future research. Furthermore, clear theoretical assumptions should be made that guide the selection of economic and cultural indicators. An interesting example in this regard is the study by Nosek et al. (2009), which assumed that in countries with pronounced stereotypical beliefs about natural sciences being a more male domain, gender gaps in mathematics and science achievement were more in favor in boys than if this was not the case. In order to investigate this association, they combined TIMSS data with results of an international implicit association test study.

Of course, the possibilities to identify actual causes of between-country differences in gender gaps in achievement with cross-sectional data from the comparative large-scale studies are very limited. Future research should reflect clearly on possible explanations for correlative findings, including selection, third variable, and reverse causality effects.

Furthermore, we would like to acknowledge that our literature review might have missed studies that did not use our search words in the title, abstract, or keywords, or that were not published in peer-reviewed journals. Therefore, an even more comprehensive literature review might reveal more interesting studies.

Embedding the Findings in a Broader Societal and Historical Context

In some countries, much has been done to counteract gender differences in access to and results of education (e.g., UNESCO, 2019). In this chapter, we focused on gender gaps in student achievement in different academic domains. While the gender gaps were close to zero in many countries and domains, partly large gender gaps either in favor of boys and, more commonly, in favor of girls are found.

Halpern's (2012) hope that international comparative studies can help to understand relations between the active factors, biology and different levels of social systems, is partly fulfilled. There seem to be few differences between boys' and girls' achievement that can be observed across all countries and that are therefore conform to stable, innate gender differences. However, there are still many questions.

Why does gender still play a role in societies? Ability and related economic incentives do not, for example, seem to affect career choices in the same way for girls as for boys (e.g., OECD, 2009). Access to birth control, preschools and schools as well as more equal division of labor between fathers and mothers in relation to children have made opportunities to make individual life-choices more equal for men and women. Such opportunities are, however, not identical and it may be that different linkages to reproduction create differences in values and in what is viewed as interesting or 'feasible' (Brownhill et al., 2015). It could be added here that the barrier for men to do 'women's work' (e.g., preschool teacher) is more difficult to break than the barrier for women to do 'men's work' (e.g., engineer). Cultural and social circumstances may thus result in subtle individual differences in attitudes and expectations. This may not have consequences for school performance but rather for the later use of abilities and knowledge (OECD, 2009). Cultural patterns and beliefs created around the core of gender divisions may enlarge small differences and keep them alive.

Why do countries vary so much in gender gaps in achievement? Information produced by the ILSAs can be used to develop a deeper understanding of processes in the gender systems by comparing countries. However, the huge masses of data are also difficult to utilize when it comes to the details. We can speculate and form complicated theoretical patterns at a general level, but when it comes to empirical testing, the complexity can become an issue. There are infinite possibilities of cultural representations and images of gender and its implications. This does not, however, mean that comparisons are uninteresting or unproductive, but the implications for action are seldom clear. The levels of gender equality norms and practices in countries (sometimes indicated by measures as the United Nation's Gender Inequality Index) are an obvious candidate as explanation for differences in gender differences, and several studies show correlations (e.g., Reilly, 2012; Rodríguez-Planas & Nollenberger, 2018). However, the links between structural gender inequalities and assessed academic achievement scores of boys and girl may not be simple. Girls are, for example, outperforming boys in reading, mathematics, and science in several Arab countries, which have a high Gender Inequality Index, that is, pronounced societal gender inequalities (UNDP, 2019). At the same time, Scandinavian countries frequently show pronounced gender gaps in favor of girls but low Gender Inequality Index values (ibid.). This shows that simple unidirectional and linear associations might not depict the whole picture. If factors that look identical on the surface interact differently with variations in the overall social structures, complex analyses are needed to make clear descriptions and explain local patterns compared to global ones. Strong theoretical foundations are needed to guide such empirical undertakings. The increasing power of computers and intricate statistical tools can help a part of the way, but we have not yet reached an evidence-based understanding of the mechanisms behind different gender systems at individual, local, country, and global levels. It is important that global generalizations may be misleading when it comes to specific systems.

The political dimension is another barrier on the way to understand gender gap mechanisms. The study of sex or gender differences in cognitive abilities and school achievement has a history of controversies and conflicts that relate to politics and power struggles. In history, major efforts have been made to prove a natural intellectual superiority of men (e.g. Shields, 1982). The everyday experiences probably supported that idea because, in most contexts before the middle of the 20th century, cognitive and academic development and freedom of action was more restricted for women than for men. In many regions of the world, societies strive to gender equality (e.g., UNESCO, 2019; World Economic Forum, 2019), but gender differences are still a debated topic. One might ask if the history of using science to prove the legitimacy of male supremacy still lingers in how problems are described, and which results are being stressed. However, an implication of what new knowledge reveals about natural or biological versus cultural or social differences is that the few small cognitive ability differences between males and females (Halpern, 2012) cannot explain the partly large gender gaps in academic achievement. The brain is complex, flexible, and prone to change in accordance with physical and social environments (cf. Halpern, 2012; OECD, 2009). Other social factors and mechanisms than cognitive ability, educational opportunity and interest have implications for how the gender structures are organized (cf. Collins et al., 1993; Saltzman Chafetz, 2006). Gender structures are complex. Collins et al. (1993), when aiming to formulate an integrated theory of gender stratification, concluded how difficult it is to construct such a theory. Some of the difficulties they encountered were:

> "[….] that evidence has come in a great variety of forms: quantitative analyses of demographic data and of surveys of attitudes, experiences, and standardized test scores; qualitative descriptions of patterns of language and interaction; historical analysis of literary sources; comparisons of economic, military, and kinship structures; studies of the dynamics of social movements. There is no simple method of reducing all this material to a comparable set of quantitative indicators which could enter into a comprehensive test of the explanatory power of the different variables. The various processes, individual characteristics, and social structures are not on the same level of analysis; the theory of gender stratification exemplifies all the problems of relating micro, meso, and macro levels, and of supplying static cross-sectional data with dynamic explanations" (Collins et al., 1993, p. 186).

This was said in 1993 and since then computer power and statistical methods have developed and it should be possible to move somewhat further towards an empirically grounded theory today. Social systems are, however, open, complex systems and change continuously. This does not, however, mean that large-scale studies, or any other type of studies, are in vain. The knowledge produced may not be complete, but it can help to understand associations between educational and societal inputs and outputs and shed light on the factors that may explain gender differences at a global level and be levers for change.

F. Conclusion

We will end this chapter by highlighting our main impressions and reflections from the review of studies on gender differences in educational achievement in ILSAs.

- First of all, both the IEA and OECD have—in all their assessments over time—produced tables that compare boys' and girls' performances on all achievement scales and subscales. Why these comparisons are important or interesting is however not explicated in any of the frameworks of the studies, nor do the reports offer any interpretations of the observed gaps or hypotheses for changes in the gaps over time. This lack of theoretical foundation may perhaps be due to traditions of educational assessments, but this needs to change. There are many good reasons to report gender differences and they should be made explicit.
- Gender differences in means and variances in educational achievement measures are usually very small compared to general inter-individual differences. The overlap between the score distributions of boys and girls most often amounts to more than 98 %.
- Gender differences in achievement are not universal, they emerge and disappear in different cultural settings. The malleability of the gaps is clearly demonstrated by the variation in the size and direction of gender differences, and the fact that they vary across countries, age groups and time.
- Some gender patterns are more stable than others. Females achieve on average higher scores than males in reading literacy, civics education and in computer and information literacy, although the degree of these gaps varies across countries.
- The gender gap trends in mathematics and science, subject domains where males traditionally have been found to achieve higher scores than females, have either decreased or disappeared over the years, and in some cases switched direction. Notably, these changes are slow. These differential trends call for explanations.
- Secondary analyses with the primary aim to deepen our understanding of gender patterns in ILSAs are scarce. Possible reasons for this may be manifold, but the lack of theory, the lack of good indicators, and the scarceness of methodological competences needed to address these phenomena with a more multivariate approach are a few examples.
- Secondary analyses of gender differences in ILSAs have often suggested non-cognitive factors such as motivation, self-concept, and self-efficacy to explain gender differences in achievement, but the empirical evidences are few and unclear. Explanatory research that includes chains of direct and indirect effects, reciprocal effects and that addresses the risk of reversed causality are still at large missing.
- ILSAs can provide unique possibilities for the analysis of group differences as the data are longitudinal at the country level. Such studies that benefit from the analysis of between-country variation are however rare, as well.
- Although gender gaps in achievement are small at both the national and international level, it should be remembered that this pattern may not be true at lower levels of observations or in different sub-groups of within-country populations. Also, rich

Western countries dominate the country samples of the ILSAs, why other gender patterns may emerge if countries from other parts of the world join.

- More focused research using ILSAs is needed that addresses gaps at different levels of observation, gap differences between countries, and changes of gender gaps across time. Such research would also benefit from including gender-related theories and from identifying factors and relations that may account for the varying gender gaps across countries and time.

In conclusion, we identified many stable results, and conceivable patterns of change. It is, however, far from fully understood how the complex web of structural mechanisms, sociocultural prerequisites and individual actions that form gender systems, produce gaps between boys' and girls' academic achievement.

G. References

Ababneh, E. G., & Abdel Samad, M. M. (2018). Gender gap in science achievement for Jordanian students in PISA2015. *European Journal of Educational Research*, *7*(4), 963–972.

Ariès, P. (1962). *Centuries of childhood: A social history of family life*. Vintage Books.

Arnot, M., David, M., & Weiner, G. (1999). *Closing the gender gap: Postwar education and social change*. Blackwell Publishers.

Ayalon, H., & Livneh, I. (2013). Educational standardization and gender differences in mathematics achievement: A comparative study. *Social Science Research*, *42*(2), 432–445. https://doi.org/10.1016/j.ssresearch.2012.10.001

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

Borgonovi, F. (2016). Video gaming and gender differences in digital and printed reading performance among 15-year-olds students in 26 countries. *Journal of Adolescence*, *48*, 45–61. https://doi.org/10.1016/j.adolescence.2016.01.004

Brownhill, S., Warin, J., & Wernersson, I. (2015). *Men, masculinities and teaching in early childhood education: International perspectives on gender and care*. Routledge. https://doi.org/10.4324/9781315756936

Chamberlain, M. C. (2003). The relative performance of girls and boys in science at grades 4 and 8 in New Zealand in 1994 and 1998. *Studies in Educational Evaluation*, *29*(3), 259–277. https://doi.org/10.1016/S0191-491X(03)90013-7

Cheema, J. R., & Galluzzo, G. (2013). Analyzing the gender gap in math achievement: Evidence from a large-scale US sample. *Research in Education*, *90*(1), 98–112. https://doi.org/10.7227/RIE.90.1.7

Cheung, K., Mak, S., & Sit, P. (2013). Online reading activities and ICT use as mediating variables in explaining the gender difference in digital reading literacy: Comparing Hong Kong and Korea. *Asia-Pacific Education Researcher*, *22*(4), 709–720. http://dx.doi.org/10.1007/s40299-013-0077-x

Chiu, M.-S. (2010). Effects of science interest and environmental responsibility on science aspiration and achievement: Gender differences and cultural supports. *Educational Research and Evaluation*, *16*(4), 345–370. http://www.informaworld.com/openurl?genre=article&id=doi:10.1080/13803611.2010.527756

Collins, R., Chafetz, J. S., Blumberg, R. L., Coltrane, S., & Turner, J. H. (1993). Toward an integrated theory of gender stratification. *Sociological Perspectives*, *36*(3), 185–216. https://doi.org/10.2307/1389242

Comber, L. C., & Keeves, J. P. (1973). *International studies in evaluation 1: Science education in nineteen countries: An empirical study*. Almqvist & Wiksell.

Connell, R. W. (2002). *Gender*. Polity Press.

De San Roman, Ainara Gonzalez, & La Rica, S. de (2016). Gender gaps in PISA test scores: The impact of social norms and the mother's transmission of role attitudes. *Estudios De Economia Aplicada*, *34*(1), 79–107.

Dronkers, J., & Kornder, N. (2015). Can gender differences in educational performance of 15-year-old migrant pupils be explained by societal gender equality in origin and destination countries? *Compare: A Journal of Comparative and International Education*, *45*(4), 610–634. http://dx.doi.org/10.1080/03057925.2014.911658

Else-Quest, N. M., Hyde, J. S., & Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin*, *136*(1), 103–127. https://doi.org/10.1037/a0018053

Fraillon, J., Ainley, J., Schulz, W., Friedman, T., & Duckworth, D. (2020). *Preparing for life in a digital world: IEA International Computer and Information Literacy Study 2018 international report*. Springer International Publishing.

Fraillon, J., Ainley, J., Schulz, W., Friedman, T., & Gebhardt, E. (2014). *Preparing for life in a digital age: The IEA International Computer and Information Literacy Study international report*. Springer International Publishing. https://doi.org/10.1007/978-3-319-14222-7

Gebhardt, E., Thomson, S., Ainley, J., & Hillman, K. (2019). *Gender differences in computer and information literacy: An in-depth analysis of data from ICILS*. Springer International Publishing. https://doi.org/10.1007/978-3-030-26203-7

Ghasemi, E., Burley, H., & Safadel, P. (2019). Gender differences in general achievement in mathematics: An international study. *New Waves - Educational Research & Development*, *22*(1), 27–54.

Gray, H., Lyth, A., McKenna, C., Stothard, S., Tymms, P., & Copping, L. (2019). Sex differences in variability across nations in reading, mathematics and science: A meta-analytic extension of Baye and Monseur (2016). *Large-Scale Assessments in Education*, *7*. http://dx.doi.org/10.1186/s40536-019-0070-9

Guiso, L., Monte, F., Sapienza, P., & Zingales, L. (2008). Diversity. Culture, gender, and math. *Science*, *320*(5880), 1164–1165. https://doi.org/10.1126/science.1154094

Halpern, D. F. (2012). *Sex differences in cognitive abilities* (4th ed.). Psychology Press.

Hamamura, T. (2012). Power distance predicts gender differences in math performance across societies. *Social Psychological and Personality Science*, *3*(5), 545–548. https://doi.org/10.1177/1948550611429191

Hermann, Z., & Kopasz, M. (2019). Educational policies and the gender gap in test scores: a cross-country analysis. *Research Papers in Education*, 1–22. https://doi.org/10.1080/02671522.2019.1678065

Hosenfeld, I., Köller, O., & Baumert, J. (1999). Why sex differences in mathematics achievement disappear in German secondary schools: a reanalysis of the German TIMSS-data. *Studies in Educational Evaluation*, *25*(2), 143–161. https://doi.org/10.1016/S0191-491X(99)00018-8

Hyde, J. S. (2005). The gender similarities hypothesis. *The American Psychologist*, *60*(6), 581–592. https://doi.org/10.1037/0003-066X.60.6.581

Hyde, J. S. (2014). Gender similarities and differences. *Annual Review of Psychology*, *65*, 373–398. https://doi.org/10.1146/annurev-psych-010213-115057

Ireson, G. (2017). Gender achievement and social, political and economic equality: a European perspective. *Educational Studies*, *43*(1), 40–50. https://doi.org/10.1080/03055698.2016.1237868

Keeves, J. P. (1992). *Learning science in a changing world: Cross national study of science achievement*. IEA Office Australia.

Kim, D. H., & Law, H. (2012). Gender gap in maths test scores in South Korea and Hong Kong: Role of family background and single-sex schooling. *International Journal of Educational Development*, *32*(1), 92–103. https://doi.org/10.1016/j.ijedudev.2011.02.009

Klieme, E. (2016). *TIMSS 2015 and PISA 2015: How are they related on the country level?* German Institute for International Educational Research (DIPF).

Maccoby, E. E. (1998). *The two sexes: Growing up apart, coming together*. Belknap Press.

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

Machin, S., & Pekkarinen, T. (2008). Assessment. Global sex differences in test score variability. *Science*, *322*(5906), 1331–1332. https://doi.org/10.1126/science.1162573

Mak, S., Cheung, K., Soh, K., Sit, P., & Ieong, M. (2017). An examination of student- and across-level mediation mechanisms accounting for gender differences in reading performance: a multilevel analysis of reading engagement. *Educational Psychology*, *37*(10), 1206–1221. https://doi.org/10.1080/01443410.2016.1242712

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. https://doi.org/10.1080/03054980701565279

Martin, M. O., Mullis, I. V. S., Foy, P., & Hooper, M. (2016). *TIMSS 2015 international results in science*. http://timssandpirls.bc.edu/timss2015/international-results/

Mead, M. (1935). *Sex and temperament in three primitive societies*. W. Morrow & Company.

Meelissen, M., & Luyten, H. (2008). The Dutch gender gap in mathematics: Small for achievement, substantial for beliefs and attitudes. *Studies in Educational Evaluation*, *34*(2), 82–93. https://doi.org/10.1016/j.stueduc.2008.04.004

Meinck, S., & Brese, F. (2019). Trends in gender gaps: Using 20 years of evidence from TIMSS. *Large-Scale Assessments in Education*, *7*. https://doi.org/10.1186/s40536-019-0076-3

Mullis, I. V. S., & Martin, M. O. (2013). *TIMSS 2015 assessment frameworks*. http://timssandpirls.bc.edu/timss2015/frameworks.html

Mullis, I. V. S., & Martin, M. O. (2015). *PIRLS 2016 assessment framework*. http://timssandpirls.bc.edu/pirls2016/framework.html

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 & PIRLS International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O., Foy, P., & Drucker, K. T. (2012). *PIRLS 2011 international results in reading*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and the International Association for the Evaluation of Educational Achievement (IEA).

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2016). *TIMSS 2015 international results in mathematics*. http://timssandpirls.bc.edu/timss2015/international-results/

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2017a). *ePIRLS 2016: The international results in online informational reading*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and the International Association for the Evaluation of Educational Achievement (IEA).

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2017b). *PIRLS 2016 international results in reading*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and the International Association for the Evaluation of Educational Achievement (IEA).

Mullis, I. V. S., Martin, M. O., Foy, P., Kelly, D. L., & Fishbein, B. (2020). *TIMSS 2019 international results in mathematics and science*. https://timssandpirls.bc.edu/timss2019/international-results/

Mullis, I. V. S., Martin, M. O., & Gonzalez, E. J. (2004). *International achievement in the processes of reading comprehension: Results from PIRLS 2001 in 35 countries*. TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O., Kennedy, A. M., & Foy, P. (2007). *IEA's Progress in International Reading Literacy Study in primary school in 40 countries*. TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O., & Loveless, T. (2016). *20 years of TIMSS: International trends in mathematics and science achievement, curriculum, and instruction*. TIMSS & PIRLS International Study Center, Boston College.

Munir, F., & Winter-Ebmer, R. (2018). Decomposing international gender test score differences. *Journal for Labour Market Research*, *52*(1). https://doi.org/10.1186/s12651-018-0246-8

Neugebauer, M., Helbig, M., & Landmann, A. (2011). Unmasking the myth of the same-sex teacher advantage. *European Sociological Review*, *27*(5), 669–689. https://doi.org/10.1093/esr/jcq038

Nosek, B. A., Smyth, F. L., Sriram, N., Lindner, N. M., Devos, T., Ayala, A., Bar-Anan, Y., Bergh, R., Cai, H., Gonsalkorale, K., Kesebir, S., Maliszewski, N., Neto, F., Olli, E., Park, J., Schnabel, K., Shiomura, K., Tulbure, B. T., Wiers, R. W., . . . Greenwald, A. G. (2009). National differences in gender-science stereotypes predict national sex differences in science and math achievement. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(26), 10593–10597. https://doi.org/10.1073/pnas.0809921106

OECD. (2001). *Knowledge and skills for life: First results from PISA 2000*. OECD Publishing.

OECD. (2002). *Reading for change: Performance and engagement across countries: Results from PISA 2000*. OECD Publishing. https://doi.org/10.1787/9789264099289-en

OECD. (2003). *Literacy skills for the world of tomorrow: Further results from PISA 2000*. OECD Publishing.

OECD. (2004). *Learning for tomorrow's world: First results from PISA 2003*. OECD Publishing.

OECD. (2009). *Equally prepared for life? How 15-year-old boys and girls perform in school*. OECD Publishing.

OECD. (2010a). *PISA 2009 at a glance*. OECD Publishing.

OECD. (2010b). *PISA 2009 results: What students know and can do: Student performance in reading, mathematics and science (Volume I)*. OECD Publishing. https://doi.org/10.1787/9789264091450-en

OECD. (2015). *The ABC of gender equality in education: Aptitude, behaviour, confidence*. OECD Publishing. https://doi.org/10.1787/9789264229945-en

OECD. (2016). *PISA 2015 Results (Volume I): Excellence and Equity in Education*. OECD Publishing. https://doi.org/10.1787/9789264266490-en

OECD. (2019a). *PISA 2018 assessment and analytical framework*. OECD Publishing. https://doi.org/10.1787/b25efab8-en

OECD. (2019b). *PISA 2018 results (volume II): Where all students can succeed*. OECD Publishing.

Penner, A. M. (2008). Gender differences in extreme mathematical achievement: An international perspective on biological and social factors. *AJS; American Journal of Sociology*, *114 Suppl*, S138-70. https://doi.org/10.1086/589252

Pereira, M. F., Fraile, M., & Rubal, M. (2015). Young and gapped? Political knowledge of girls and boys in Europe. *Political Research Quarterly*, *68*(1), 63–76. https://doi.org/10.1177/1065912914554040

Postlethwaite, T. N., & Wiley, D. E. (1992). *The IEA study of science II: Science achievement in twenty-three countries (Volume 2)*. Pergamon.

Reilly, D. (2012). Gender, culture, and sex-typed cognitive abilities. *PLoS ONE*, *7*(7). https://doi.org/10.1371/journal.pone.0039904

Reilly, D., Neumann, D. L., & Andrews, G. (2019). Investigating gender differences in mathematics and science: Results from the 2011 Trends in Mathematics and Science Survey. *Research in Science Education*, *49*(1), 25–50. http://dx.doi.org/10.1007/s11165-017-9630-6

Rodríguez-Planas, N., & Nollenberger, N. (2018). Let the girls learn! It is not only about math … it's about gender social norms. *Economics of Education Review*, *62*, 230–253. https://doi.org/10.1016/j.econedurev.2017.11.006

Rosén, M. (1998). *Gender differences in patterns of knowledge. Göteborg Studies in Educational Sciences: Vol. 124*. Acta Universitatis Gothoburgensis. http://hdl.handle.net/2077/13820

Saini, A. (2017). *Inferior: How science got women wrong - and the new research that's rewriting the story*. 4th Estate.

Saltzman Chafetz, J. (Ed.). (2006). *Handbooks of Sociology and Social Research. Handbook of the sociology of gender* (1st ed. 2006). Springer US; Imprint: Springer.

Schulz, W., Ainley, J., Fraillon, J., Kerr, D., & Losito, B. (2010a). *ICCS 2009 international report: Civic knowledge, attitudes, and engagement among lower-secondary school students in 38 countries*. International Association for the Evaluation of Educational Achievement.

Schulz, W., Ainley, J., Fraillon, J., Kerr, D., & Losito, B. (2010b). *Initial findings from the IEA International Civic and Citizenship Education Study*. International Association for the Evaluation of Educational Achievement.

Schulz, W., Ainley, J., Fraillon, J., Losito, B., Agrusti, G., & Friedman, T. (2018). *Becoming citizens in a changing world: IEA International Civic and Citizenship Education Study 2016 international report*. Springer International Publishing. https://doi.org/10.1007/978-3-319-73963-2

Schwabe, F., McElvany, N., & Trendtel, M. (2015). The school age gender gap in reading achievement: Examining the influences of item format and intrinsic reading motivation. *Reading Research Quarterly*, *50*(2), 219–232. https://doi.org/10.1002/rrq.92

Shera, P. (2014). School effects, gender and socioeconomic differences in reading performance: A multilevel analysis. *International Education Studies*, *7*(11), 28–39.

Shields, S. A. (1982). The variability hypothesis: The history of a biological model of sex differences in intelligence. *Signs*, *7*(4), 769–797. http://www.jstor.org/stable/3173639

Siddiq, F., & Scherer, R. (2019). Is there a gender gap? A meta-analysis of the gender differences in students' ICT literacy. *Educational Research Review*, *27*, 205–217. https://doi.org/10.1016/j.edurev.2019.03.007

Smith, D. S., Wendt, H., & Kasper, D. (2017). Social reproduction and sex in German primary schools. *Compare: A Journal of Comparative & International Education*, *47*(2), 240–256. https://doi.org/10.1080/03057925.2016.1158643

Stoet, G., & Geary, D. C. (2013). Sex differences in mathematics and reading achievement are inversely related: Within- and across-nation assessment of 10 years of PISA data. *PLoS ONE*, *8*(3). https://doi.org/10.1371/journal.pone.0057988

Stoet, G., & Geary, D. C. (2018). The gender-equality paradox in science, technology, engineering, and mathematics education. *Psychological Science*, *29*(4), 581–593. https://doi.org/10.1177/0956797617741719

Stoet, G., & Geary, D. C. (2020). Corrigendum: The gender-equality paradox in science, technology, engineering, and mathematics education. *Psychological Science*, *31*(1), 110–111. https://doi.org/10.1177/0956797619892892

Tao, H.-L., & Michalopoulos, C. (2018). Gender equality and the gender gap in mathematics. *Journal of Biosocial Science*, *50*(2), 227–243. https://doi.org/10.1017/S0021932017000141

Torney-Purta, J., Lehmann, R., Oswald, H., & Schulz, W. (2001). *Citizenship and education in twenty-eight countries: Civic knowledge and engagement at age fourteen*. International Association for the Evaluation of Educational Achievement.

Tsai, S.-L., Smith, M. L., & Hauser, R. M. (2018). Gender gaps in student academic achievement and inequality. *Research in the Sociology of Education*, *20*, 181–218. https://doi.org/10.1108/S1479-353920180000020008

UNDP. (2019). *Human development report 2019: Beyond income, beyond averages, beyond today: Inequalities in human development in the 21st century*. Bernan.

UNESCO. (2018). *Sdg 4 data digest: Data to nurture learning*. Unesco Institute for Statistics.

UNESCO. (2019). *Building bridges for gender equality*. Unesco. https://en.unesco.org/gem-report/2019genderreport

van Hek, M., Kraaykamp, G., & Pelzer, B. (2018). Do schools affect girls' and boys' reading performance differently? A multilevel study on the gendered effects of school resources and school practices. *School Effectiveness & School Improvement*, *29*(1), 1–21. https://doi.org/10.1080/09243453.2017.1382540

van Langen, A., Bosker, R., & Dekkers, H. (2006). Exploring cross-national differences in gender gaps in education. *Educational Research & Evaluation*, *12*(2), 155–177. https://doi.org/10.1080/13803610600587016

Wernersson, I. (2010). *Könsskillnader i skolprestationer – idéer om orsaker (Gender differences in school achievement – ideas about causes)*. Statens Offentliga Utredningar. https://www.regeringen.se/49b719/contentassets/d48f0f7bf67a4d8b8b62b3dfbc36b551/konsskillnader-i-skolprestationer---ideer-om-orsaker-sou-201051

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

Wiseman, A. W. (2008). A culture of (in)equality? A cross-national study of gender parity and gender segregation in national school systems. *Research in Comparative and International Education*, *3*(2). http://dx.doi.org/10.2304/rcie.2008.3.2.179

World Economic Forum. (2019). *The global gender gap report 2020*. World Economic Forum.

Yamamura, E. (2019). Gender wage gap and its effect on test scores of immigrant students. *Journal of Economic Studies*, *46*(4), 872–887. https://doi.org/10.1108/JES-04-2018-0141

Zhou, Y., Fan, X., Wei, X., & Tai, R. (2017). Gender gap among high achievers in math and implications for STEM pipeline. *Asia-Pacific Education Researcher*, *26*(5), 259–269. https://doi.org/10.1007/s40299-017-0346-1

H. Appendix: Details of the Literature Search

Specifically, we ran a search for references that mentioned three keyword groups at the same time in the title, keywords, and/or abstract. The three keyword groups were the following: (1) “gender difference” OR “gender gap” OR “sex difference” OR “sex gap” OR “gender inequality” OR “sex inequality” OR “gender equity” OR “sex equity” OR “gender fairness” OR “sex fairness” OR “gender equality” OR “sex equality”, (2) AND “achievement” OR “attainment” OR “performance” OR “ability” OR “aptitude” OR “competence” OR “test score” OR “literacy” OR “reading” OR “mathematics” OR “science” OR “civic” OR “citizenship” OR “information” OR “computer” OR “technology” OR “digital”, (3) AND “Programme for International Student Assessment” OR “PISA” OR “Progress in International Reading Literacy Study” OR “PIRLS” OR “Trends in International Mathematics and Science Study” OR “Third International Mathematics and Science Study” OR “TIMSS” OR “Civic Education Study” OR “CIVED” OR “International Civic and Citizenship Education Study” OR “ICCS” OR “International Computer and Information Literacy Study” OR “ICILS” OR “international studies” OR “comparative analysis” OR “international assessment”. We limited the search to peer-reviewed journal articles that were published in English and between 1995 and 2019. By using EbscoHost, we searched the databanks Academic Search Premier, ERIC, PsycARTICLES, PsycINFO, PSYNDEX, SocINDEX, Education Source, Education Research Complete, and EconLit. Additionally, we searched Web of Science.