International Reading Gaps between Boys and Girls, 1970–2016

Isa Steinmann, Rolf Strietholt, and Monica Rosén

Author Note

The work is part of the Center for Comparative Analyses of Educational Achievement (COMPEAT), which is supported by the Swedish Riksbankens Jubileumsfond [grant number In2006-0401:1-IK], the research project Changes in Opportunities to Learn: Swedish Pupils Results in Large International Studies over Time, which is funded by the Swedish Research Council (Vetenskapsrådet) [grant number 2015-01080], and the European Training Network Outcomes and Causal Inference in International Comparative Assessments (OCCAM), which is supported by the European Union’s Horizon 2020 research and innovation program [Marie Skłodowska-Curie grant number 765400].

We thank Roisin Cronin for copy editing the manuscript, Simon Beesch for his assistance in the literature review, Laura Zieger and Hannah Marie Heister for their assistance with the analyses, and Thomas Kiefer for the opportunity to use the beta version of the R package TAM.

Correspondence concerning the article should be addressed to Isa Steinmann, Oslo Metropolitan University, Norway. E-mail: isa.steinmann@oslomet.no

Abstract

Gender differences are one of the most contentious issues in educational research. This study analyzes long-term changes in gender gaps in reading comprehension at the end of primary school in 63 education systems. It links test data from seven comparative studies that were conducted between 1970 and 2016 using a common achievement scale based on item response theory. We investigate whether mean gender gaps have widened or narrowed over time—controlling for changes in the sample of countries from measurement point to measurement point—using a system-level regression with fixed effects. We observe an advantage of girls over boys in reading in almost all countries, although the size of the gender gap varies considerably internationally. Further, we observe a significant increase in the international gender gap between 1971 and 2001 and a slight decrease since then.

Keywords: reading literacy, gender gap, international large-scale assessment, item-response theory, linking

International Reading Gaps between Boys and Girls, 1970–2016

The question of gender equality in reading achievement—i.e., of whether boys and girls perform differently—is a contentious issue in educational research. A student's ability to read fluently and to comprehend what they are reading is an important outcome of primary education; it has an intrinsic value when children read for pleasure and an instrumental value when they read to learn in other domains or use reading to participate in the society. More broadly, gender equality in education has also been recognized as an important issue by policy makers. For instance, in 2015, the United Nations Educational, Scientific and Cultural Organization (UNESCO) prominently stressed gender equality and quality education in two of the 17 global goals for a sustainable development (UNESCO 2016).

Theoretical Perspectives on the Emergence of Gender Gaps in Academic Achievement

Theories differ in how they understand the emergence of gender differences in academic achievement (see overviews by Maccoby 1998; Connell 2002; Halpern 2012). One set of arguments assumes that males and females had different innate cognitive abilities, that is, they differ in their stable, biological capacities to learn. However, extensive meta-analyses on this topic have found very small gender differences in most subdomains of cognitive ability tests. The only somewhat larger stable differences were found in the verbal and visual-spatial areas, with females performing slightly better in the former and males slightly better in the latter, but these were still small (e.g., Maccoby and Jacklin 1974; Rosén 1998; Halpern 2012). In Hyde's (2005) review of meta-analyses, for instance, 69% of the reported gaps between boys and girls or men and women in cognitive variables were very small or close to zero ($d < |0.20|$), 23% were small to medium ($d < |0.50|$), and only 8% were medium to large ($d < |0.80|$). The effect size d reflects mean differences between the genders with $d = 1$ implying a one standard deviation advantage of females and $d = -1$ a one standard deviation advantage of males. In other areas, much larger gender differences were observed, for instance, in throwing velocity (*d* = -2.18). Hyde (2005) concluded that “males and females are alike on most—but not all—psychological variables” (ibid, 590).

Of course, academic achievement is not the product of innate cognitive abilities alone but also of motivational and learning processes that are strongly tied to the environment. Many theoretical perspectives assume that important environmental factors differ for boys and girls. For example, multiple gender stereotypes and gender-specific expectations are assumed to influence child development. This means that boys and girls are positively and negatively reinforced to show gender-appropriate behavior on a daily basis, for example, by their families, peers, teachers, or the media (Connell 2002; Halpern 2012). Following Maccoby (1998), “there can be no doubt that as the cultural messages, the social assumptions, and the scripts concerning gender are absorbed by children, they have a powerful impact on the way children construct their identity as either male or female individuals” (ibid, 294).

According to reading-related gender stereotypes, girls are better at language-related tasks and more interested in reading. Such stereotypes have two clear implications: They could, on the one hand, prompt boys to develop lower reading-related self-esteem and test scores due to stereotype threat mechanisms (e.g., Retelsdorf et al. 2015; Wolter et al. 2015; Pansu et al. 2016). On the other hand, they could contribute to less engagement in reading as a hobby among boys (e.g., expectancy-value theory; Wigfield and Eccles 2000). Indeed, boys typically show lower reading self-esteem, lower self-efficacy in reading, and more negative attitudes towards reading than girls (e.g., Archambault et al. 2010; Huang 2013; OECD 2015). Importantly, and arguably related to the aforementioned findings, boys read less often in their free time than girls, especially fiction (e.g., OECD 2009; OECD 2015; Jerrim and Moss 2019). Since reading enjoyment and regular reading outside of school are regarded as important promoters of reading achievement (e.g., OECD 2015; Waxman 2015; Jerrim and

Moss 2019), such gender differences in reading habits might influence gender gaps in reading performance.

Stereotypes or "cultural messages" (Maccoby 1998, 294) about girls and boys' academic performances likely differ between countries and change in line with social, political, and economic developments (e.g., Connell 2002; Assié-Lumumba and Sutton 2004; Cooray and Potrafke 2011). Different previous studies correlated countries' gender gaps in achievement with indicators of societal gender inequalities, assuming relative female achievement advantages in more gender-equal societies. The indicators differed however vastly between the studies, including gender-related attitudes from the World Values Survey, the World Economic Forum's Gender Gap Index, female enrolment ratios in specific university tracks, or gender gaps in salaries or labor force participation. Some of these studies indeed found gender gaps to be more shifted in favor of girls in more gender-equal countries (e.g., Guiso et al. 2008; Reilly 2012; van Langen et al. 2006), others, however, found mixed or insignificant associations (e.g., Else-Quest et al. 2010; Marks 2008; Stoet and Geary 2013; Tao and Michalopoulos 2018). This inconclusive state of research may be related to the use of different indicators and country samples. A study that used a direct measure for relevant societal gender stereotypes, namely societies' implicit stereotypes of science being a male domain, found a close link with gender achievement gaps (Nosek et al. 2009).

Apart from gender-related cultural aspects, countries' educational policies have the potential to affect gender gaps in student achievement, especially if the policies lead to boys and girls being segregated into different schools. According to the opportunity to learn theory (McDonnell 1995), the potential for gender gaps in learning processes should be larger in systems where boys and girls are segregated in different schools with potentially different qualities. Van Langen and colleagues (2006) found that differentiation policies indeed correlated with gender gaps in achievement. They constructed a country-level indicator of overall differentiation in the education system, including single-sex schooling and tracking policies, among others. Gender gaps were more shifted in favor of girls in more integrated as compared to differentiated systems (van Langen et al. 2006). In contrast, Hermann and Kopasz (2019) compared countries with tracked and compulsory lower secondary school systems and found girls' advantages in the first as compared with the latter. Other studies that investigated single-sex versus coeducational schooling within the same countries found no robust effects on achievement outcomes (e.g., Pahlke et al. 2014; Robinson et al. 2021). Therefore, the state of research on differentiation policies is inconclusive.

Another country characteristic that has the potential to directly affect achievement gender gaps is the school enrolment ratio. Especially at the secondary school level, in lower-income areas of the world or a few decades ago, not all children and adolescents attended school (e.g., OECD 2015; UNESCO 2019). By tendency, enrolment gaps are nowadays shaped to the advantage of boys in lower- and to the advantage of girls in higher-income areas of the world (e.g., UNDP 2019; World Economic Forum 2019). A low socioeconomic status is, internationally, one of the most central risk factors for not being enrolled in school (e.g., Lavrijsen and Nicaise 2015; OECD 2020; UNDP 2019). If either the enrolment of boys or girls increases in a country, more socioeconomically disadvantaged and probably low-performing students enter or remain in school, which should therefore affect gender gaps in school-based assessments. Indeed, Steinmann and Rutkowski (in print) found a negative association between countries' gender gaps in school enrolment and gender gaps in academic achievement scores at the secondary school level.

In summary, various differences between countries, such as cultural, political, and school enrollment differences, might explain why gender gaps in achievement could vary between countries and over time. Based on the presented literature, we expected to find that girls achieve higher reading scores on average than boys and that these gender gaps differ between countries and across time in the present study.

Evidence on Gender Gaps in Reading Comprehension

Many believe that girls generally score higher than boys in reading, but previous research provides a more nuanced picture. We start our literature review with Hyde and Linn's (1988) meta-analysis, which summarized 165 US and Canadian studies on gender differences in verbal abilities that were conducted between 1955 and 1986. The main finding was that girls had better verbal ability scores although the mean effect size was small (d = 0.11).[1] The meta-analysis revealed some further interesting results. First, the gender gaps varied for the subdomains of verbal ability. The gender differences ranged from d = -0.16 in analogies, meaning that boys had a very small advantage, to d = 0.33 in speech production, implying that girls had a small to moderate advantage. The mean effect size for reading comprehension was close to zero (d = 0.03). Second, the authors found some evidence that gender gaps declined slightly over time, as the gaps found in studies published in 1973 or earlier (d = 0.23) were larger than those found in studies published after 1973 (d = 0.10). Third, the mean effect sizes were similar for children of different age groups.

Another meta-analysis on gender gaps summarized international and national large-scale assessments on reading comprehension at secondary school level conducted between 1970 and 2002 (Lietz 2006a; Lietz 2006b). The international data stemmed from around 50 countries that participated in the international Reading Comprehension Study (RCS) 1970, the Reading Literacy Study (RLS) 1991, and the Programme for International Student

[1] For our literature review, we transformed all original mean difference scores into effect size measures to increase the comparability of the findings. If no effect size measures were reported, we divided the mean score differences of girls minus boys by the (international) standard deviation. Therefore, all d values can be interpreted as differences in standard deviation units, with positive values indicating an average advantage for girls and negative values indicating an average advantage for boys.

Assessment (PISA) 2000, as well as from the national data from the National Assessment of Educational Progress (NAEP) 1992, 1994, 1998, 2002, and 2003 in the US and the Australian Monitoring Standards in Education (MSE) 1992, 1995, 1997, 1999, 2001, and 2002. The meta-analysis treated each participating country and year of data collection as a separate study. Hence, there were a total of 147 studies, with each containing representative data from several thousand students. Girls had higher scores than boys in most studies; the mean effect size across all countries and over time was $d = 0.19$. Again, there was some evidence of variation in the size of gender gaps. In the first paper based on the meta-analysis, Lietz (2006b) showed that gaps were very small in studies conducted in 1991 or earlier ($d = 0.05$) and more pronounced in studies conducted after 1991 ($d = 0.31$).[2] However, in a second paper on the same meta-analysis, Lietz (2006a) also noted differences in the reading comprehension tests between assessment programs. PISA, NAEP, and MSE reported larger gender differences than the other assessments. It is impossible to decide whether the observed differences are due to increasing gaps or differences in study characteristics, such as the reading comprehension tests used or the country compositions in the various studies. The meta-analysis provided no evidence of age-related differences in gender gaps. As it only included secondary school studies, the variation in student age was, however, small.

Lynn and Mikk (2009) summarized the findings of two international assessments on reading comprehension at different educational stages. They considered primary school data from the Progress in International Reading Literacy Study (PIRLS) 2001 and 2006 and secondary school data from PISA 2000, 2003, and 2006. Both assessment programs thus covered approximately the same short period. Lynn and Mikk (2009) found that girls scored higher than boys, with a mean effect size of $d = 0.23$ in primary schools and a mean effect

[2] The effect size estimates for studies published before and after 1991 were based on regression analyses (studies as cases) where the gender gap was regressed on a dichotomous time variable.

size of *d* = 0.42 in secondary schools. Again, it should be noted that PISA and PIRLS were conducted in different countries and used different achievement tests, which could perhaps explain the higher effect sizes at the secondary school level.

This literature review has suggested that, internationally, reading gender gaps vary: Some studies report differences close to zero, while others find moderate advantages for girls. However, the comparability of the different meta-analyses is limited because the primary studies investigated different outcomes, employed different test instruments, and sampled from different target populations. Such meta-analyses assume that variation across outcomes, tests, and samples is comparable and thus use the outcome variation in the respective study samples to standardize and merge the observed effects across studies and time. However, researchers wishing to grasp actual national or international trends in a certain effect need outcome operationalizations, test instruments, and samples to be more consistent. This consistency can be found in national and international trend studies, which employ the same assessment and sampling frameworks across study cycles.

One such long-term trend study with comparable instruments and samples is NAEP, which has assessed grade 4, 8, and 12 in the United States since 1971. Hedges and Nowell (1995) focused on the grade 12 trend samples that were assessed between 1971 and 1992. In all cycles, girls had higher reading comprehension scores than boys, with effect sizes between *d* = 0.18 and *d* = 0.30. There was no apparent trend in the varying effect sizes over time. Klecker (2006) studied more recent data from all NAEP cohorts between 1992 and 2003 and found that girls had significant advantages in reading in all age groups and across all years of analysis. For grade 4 students, the effect sizes ranged from *d* = 0.13 to *d* = 0.27 across the years of observation, and for grade 8 students, they ranged from *d* = 0.27 to *d* = 0.43. Again, there was no clear trend over time. In 12th-graders, the gap increased from *d* = 0.22 in 1992 to *d* = 0.44 in 2002 (see also alternative study by Waxman 2015).

The Early Childhood Longitudinal Study (ECLS) is a US panel study. Based on ECLS, Chatterji (2006) analyzed changes in language and reading achievement gender gaps from kindergarten to grade 1 in more than 2,000 children. It should be noted that the reading test measured reading comprehension as well as basic reading skills such as letter recognition. Chatterji found that girls enjoyed increasing and significant advantages across time. The size of the gap increased from $d = 0.17$ at kindergarten entry to $d = 0.31$ at the end of first grade when controlling for ethnicity and poverty.

These national studies have provided interesting insights concerning trends in the United States but—as discussed above—international and cultural differences are likewise plausible. Therefore, it is worth further reviewing international evidence from comparative studies (cf. Hanushek and Woessmann 2011).

Trends in Gender Gaps in International Large-Scale Assessments

The International Association for the Evaluation of Educational Achievement (IEA) has conducted international large-scale assessments on reading literacy among schoolchildren since 1970 in an ever-increasing number of countries. In some cases, these studies were also included in the meta-analyses above. Therefore, in this section, we focus on additional trend findings that they provide. The first IEA study on reading abilities was the RCS from 1970 (Thorndike 1973). Since this study was not repeated with the same sampling and testing procedures, trend analyses with later findings cannot be directly undertaken. The next IEA study on reading was the RLS, which surveyed reading achievement and reading activities from 32 countries in 1990/1991 (Elley 1992; Raudenbush et al. 1994). In nine of these countries, the RLS was repeated in 2001 using common measurement metrics and sampling designs. Martin et al. (2003) investigated the reading gender gap changes in these nine countries and found that the gap in favor of girls significantly increased in Singapore and decreased to the point of being insignificant in Iceland and Italy. In the other six countries

(Greece, Hungary, New Zealand, Slovenia, Sweden, and the United States), girls retained an approximately similar-sized advantage over boys over the 10 years.

Since 2001, the IEA has conducted PIRLS in a five-year cycle. Because the instruments and samples are comparable in the consecutive cycles, reading gender gap trends can be observed over 15 years. The most recent report contains an overview of gender gap trends in the 49 countries and benchmarked participants that took part in at least two PIRLS cycles (TIMSS & PIRLS International Study Center 2017). In 26 of these countries, the changes between the earliest and latest time of participation were very small ($d < 0.05$). In nine countries, the gaps increased and in 14, they decreased. The reading gap only increased from an insignificant advantage of girls to a significant one in two countries (Israel and Spain) and in two other territories, the reading gap decreased from a significant advantage of girls to an insignificant advantage over time (Andalusia in Spain and Portugal). Looking at the countries that participated in all four PIRLS assessments, one can mostly see largely stable (France, Germany, Hong Kong, Hungary, Italy, Lithuania, Norway, Russian Federation, Slovenia) and decreasing gender gaps (Bulgaria, England, Netherlands, New Zealand, Singapore, Slovak Republic, Sweden, United States), instead of increasing gaps (Iran).

For adolescents, PISA—conducted by the Organization for Economic Co-Operation and Development (OECD) in a three-year cycle since 2000—is the broadest international large-scale assessment that investigates reading literacy. Trend analyses between PISA 2000 and 2006 and between 2009 and 2015 indicated slight declines in gender gaps in reading internationally (OECD 2009; OECD 2016; OECD 2019). Yet such changes varied in degree between countries, and in many countries, the gender gaps remained stable, just as found in PIRLS.

The Survey of Adult Skills (PIAAC)—conducted by the OECD between 2011 and 2018—found that, in most countries, the gender gaps between men and women in literacy skills were small and not significantly different from zero (OECD 2019). Comparisons of adult cohorts by age showed that gender gaps tended to favor women in younger groups and men in older groups. However, the observed reading advantages of young women in PIAAC were less pronounced than the reading advantages of female 15-year-olds in PISA (Borgonovi et al. 2017).

In summary, studies that used older data found mostly small average gender gaps in reading in meta-analyses and in national and international large-scale studies. The estimates varied depending on the investigated countries and study characteristics. The more recent international large-scale studies—PIRLS and PISA—found that females had significantly higher scores than males in reading across almost all countries and study cycles. The magnitude of the gender gap not only varied across time but also between countries. However, it is important to note that the more recent studies not only consider a larger set of countries but also countries from more regions of the world. For this reason, it is difficult to disentangle if the changes in the gap size reflect an actual trend in the gaps or rather the fact that the estimations base on different countries. Short-term trend analyses showed a heterogeneous picture when comparing countries. To date, no studies have simultaneously investigated differences between educational systems and longer-term trends in reading gender gaps.

The Present Study: Investigating International Gender Gaps in Reading between 1970 and 2016

The present study mainly aimed to investigate how international reading gender gaps in primary school students have changed since 1970. The above-cited literature underscores the need to identify the scope of a study on long-term international trends carefully in three regards to eliminate related bias. First, it is important to focus on individual educational stages. In the present study, we investigated gender gaps at the end of primary school (grades 3–6). Second, the observed samples should be coherent over time. We analyzed recent and older international assessments and included all observed countries that participated in at least two years of observation to be able to measure change. In order to control for changes in the sample composition over time, we used country-level regression models with fixed effects for countries. Third, it is important to use comparable measures across countries and time. The international assessments provide achievement tests that were equivalent across countries for different points in time. We built on previous work to further link the achievement tests from all reading assessments conducted by the IEA within a common item response theory (IRT) based scale (Strietholt and Rosén 2016). These scores provide a common metric for investigating gender gaps across countries and over time.

Methods

Sample

This study combines data from all seven IEA studies on reading comprehension at the end of primary school that have been conducted so far (see Table 1). These were the RCS 1970, RLS 1991 and 2001, and PIRLS[3] 2001, 2006, 2011, and 2016. We also included a Swedish extension of the international design, in which an additional sample of students responded to an extended set of test items.[4] We merged the samples of countries that participated in both RLS and PIRLS in 2001 (cf. Strietholt et al. 2013) and the samples of

[3] We did not include data from the PIRLS Literacy, prePIRLS, or ePIRLS studies.

[4] Together, all used datasets are available from https://timssandpirls.bc.edu/pirls-landing.html and https://www.gu.se/en/compeat. Sweden extended the international assessment design by including more reading items that were used in earlier studies in RLS 1991 and 2001, as well as PIRLS 2001 (see Appendix A2).

countries that assessed more than one grade in a cycle. After excluding countries or regions[5] that participated in only one study—for which changes over time could thus not be investigated—we reduced the original full sample of *n* = 234 country-by-year observations (column 1 in Table 1) to *n* = 213 (column 2 in Table 1). Out of the 62 included countries, four participated in all seven studies, eight in at least six of the studies, 14 in at least five, 27 in at least four, and 45 in at least three.

< Table 1 about here >

Within countries, we excluded students who did not take the reading test[6] or for whom gender information was missing[7]. Across the 213 country-by-year observations, we used data from almost one million students (column 3 in Table 1). Depending on the country-by-year observations, these students attended grades 3–6 and were on average between 8.9 and 11.9 years old.[8] Appendix A1 depicts a full list of study participation and student samples for the countries.

Instruments

Reading Comprehension

In all studies, the reading tests consisted of text passages and corresponding items (Thorndike 1973; Martin et al. 2003; Mullis et al. 2017). However, even though there were

[5] In the IEA studies, some countries had regional samples rather than nationwide ones (e.g., Belgium (Flemish) and Belgium (French)). We treated these as separate samples. In the following, we use the term “country” for the sake of simplicity.

[6] In RCS and RLS, students who did not participate in the reading tests were included in the datasets; they were not included in the PIRLS datasets.

[7] The shares of missing gender information ranged between 0% and 11%. While most countries had no or only very small proportions of missing gender data (< 3%), there were seven country-by-year observations with 3%–5% missing data, and two with 7%–11% missing data. These cases primarily occurred in the older studies, RCS 1970 and RLS 1991.

[8] The international target populations in RCS were the grades with most 10-year-old students, in RLS they were the grades with most 9-year-old students, and in PIRLS students they were these in grade 4. The RCS 1970 sample included some students from grades 7–13. We excluded those from the analyses. In PIRLS, some countries sampled grades above and below grade 4 (e.g., if students in grade 4 were on average not fluent readers yet or younger than 9.5 years). In some cases, the grade information from the raw datasets deviated from the information in the international reports. In this case, we used the information from the international reports (see Appendix A1).

overlaps in the assessment material, the scores in RCS, RLS, and PIRLS are not comparable over time. To establish a common scale for all assessments, we conducted a test equation study, which is explained in detail elsewhere (Strietholt and Rosén 2016). In brief, we first defined the construct of interest and identified the corresponding texts and items. We focused on reading for literary experience and to acquire and use information.[9] After reading narrative or expository text passages, students responded to multiple choice or constructed response items that assessed their comprehension of these continuous texts. In the multiple choice items, students had to choose the correct answer (1 point) out of four options. In constructed response items (i.e., free, unstructured responses to questions), students could receive between 1 to 3 points. We recoded all item responses in the assessments consistently (*wrong* or *omitted* (0), *correct, 1 point* (1), *correct, 2 points* (2), *correct, 3 points* (3), and *not presented* or *not reached* (missing)).[10] Overall, the selected test materials included 32 text passages (15 narrative and 17 expository texts) and a total of 300 corresponding items (189 multiple choice and 111 constructed response).

Second, we estimated overall reading comprehension scores on our own common metric across the selected items in the seven studies (see detailed description in Strietholt and Rosén 2016). In summary, we took advantage of the fact that many text passages and items were used in multiple assessments over time, i.e., they served as anchor items across assessments. It is important to note that in addition to the overlaps in the international design, Sweden supplemented the international tests with items from earlier tests in 1991 and 2001. These unique design features in Sweden enable the linking of all studies (see Strietholt and

[9] We excluded items that measured word recognition, so-called document items (i.e., where students had to retrieve information from tables, charts, etc.), and items from the PIRLS Reader (which were assessed in a separate booklet that did not follow the same booklet rotation principles of the other studies).

[10] RCS and the Swedish extension of RCS items in RLS 1991 and 2001, did not differentiate between omitted, not reached, and not presented items in the original datasets.

Rosén 2016 for a detailed overview). Appendix A2 gives a full overview of the text passages in the study cycles. Based on the raw data for the subsample of four countries that participated in all seven assessments (Sweden, Hungary, Italy, and United States), we estimated the item parameters in a concurrent calibration based on a Rasch model (one-parameter logistic IRT) with an extension for partial credit (for the constructed response items in which students could get 2 or 3 points; see Kim and Cohen 2002; Masters 1982).[11] We used this model's fixed item parameter estimates to estimate the person parameters for all countries. Specifically, we derived five plausible values of reading comprehension for all students in all country-by-year observations (i.e., no missing values in these plausible values).[12] We standardized each plausible value to a mean of zero with a standard deviation of one.[13] We used the R package TAM to estimate the multi-group IRT models (Kiefer et al. 2016).

Gender Gaps

As mentioned in the sample section, all students with missing gender information were excluded from the analysis. The share of girls in the samples ranged between 42% and 60% between the country-by-year observations. For each country-by-year observation, the gender gap was computed as the mean difference between the weighted[14] mean reading scores of girls and boys across five plausible values. Since we used z-standardized achievement scores, the mean difference can be interpreted as effect size d. Therefore, gender

[11] In unreported analyses, we also fitted a more complex three-parameter logistic IRT model. Due to its complexity, this model led to partly unstable results and convergence issues. The converged models resulted in similar gender gap estimates as the simple one-parameter models, which is why we decided to focus on the simple, parsimonious models.

[12] When estimating the plausible values, we only included achievement information and no additional background information.

[13] In this standardization procedure, each country-by-year observation had the same weight.

[14] We applied the student sampling weights "supwgt" (RCS 1970), "stdwgt" (RLS 1991), and "HOUWGT" (RLS 2001 and PIRLS 2001–2016). These weights account for unequal selection probabilities and non-response in the stratified clustered sampling designs in the respective studies and therefore allow to estimate gender gaps in the underlying student populations.

gaps of d = 1 imply that girls' mean achievement is one standard deviation higher than boys' mean achievement in a country, and d = -1 a one standard deviation advantage of boys.

In order to evaluate whether the results of our gender gap estimation procedures (i.e., country-by-year observation selection and reading comprehension calibration) were comparable with previously published study reports, we compared our findings with the official gender gap trends of PIRLS 2001–2016 (Mullis et al. 2017). Comparing Figure 1 and Appendix A3, we found very similar estimated within-country trends between 2001 and 2016, which provides evidence that our analyses replicate the international studies well. The effect sizes of our findings were closer to zero than the IEA's mean difference scores divided by the international standard deviation of 100, because our mean differences were evaluated against a larger overall variation in reading achievement in the country-by-year observations between 1970 and 2016.

Empirical Model

By linking the datasets and estimating the gender gap, we generated a database with one gender gap estimate for each country-by-year observation. In order to estimate change in these gender gaps, we regressed the gender *gap* in country *c* and year *t* on the variable $year$ when the assessment was administered:

$$gap_{ct} = \alpha + \beta * year_{ct} + \varepsilon_{ct} \qquad (1)$$

The key parameter of interest is β, which reflects the linear international change in the gender gap per year. The main challenge of this approach is that different countries participated in different years. If, for example, more countries with larger gender gaps participated in older studies, the comparison with more recent studies would be biased. To avoid potential bias emerging from changes in sample composition, we extended the regression model by country-fixed effects ν, i.e., we added dummies for all countries:

$$gap_{ct} = \alpha + \beta * year_{ct} + \nu_c + \varepsilon_{ct} \qquad (2)$$

The key advantage of this approach is that it exploits within-country variation to estimate gender gap trends over time. By implication, β reflects the linear change per year in the gender gap across countries, independent of the time-varying country participation.

Previous research suggests that changes in the international gender gap trends might be nonlinear (e.g., Hyde and Linn 1988; Lietz 2006b). To capture such nonlinearity, we replaced the continuous time variable with dummies for each assessment year, using 1970 as the reference. As above, we estimated country-fixed effects:

$$gap_{ct} = \alpha + \beta_1 * year_{1991c} + \beta_2 * year_{2001c} + \beta_3 * year_{2006c} + \beta_4 * year_{2011c} + \beta_5 year_{2016c} + \nu_c + \varepsilon_{ct} \qquad (3)$$

Lastly, we decomposed the variance in gender gaps in country-by-year observations into three components, one for time-stable, between-country differences, one for the overall international trend, and one for national trends (country deviations from the international trend). For this purpose, we conducted an analysis of variance with the two categorical factors, country and year, and their interaction. The interaction term reflects the country-specific changes over time. Note that including an interaction term for the country-specific trends leads to a model with zero degrees of freedom, which is why we cannot conduct a significance test.

Results

Descriptive Results

We observed positive small- to medium-sized gender gaps in almost all of the 213 country-by-year observations, i.e., girls generally had higher reading comprehension scores than boys (see Figure 1). Table 2 presents the descriptive distributions of the gender gaps separately for each year of observation. Across all country-by-year observations, the mean gender gap effect size was d = 0.14. This implies that girls scored 14% of a standard deviation higher than boys, on average. However, there was considerable variation in the size of the gaps. We only observed negative gaps, which indicated very small or close to zero advantages for boys, in three country-by-year observations (d = -0.08 in the Netherlands in 1970, d = -0.02 in Colombia in 2011, and d = -0.01 in Hungary in 1970). For 48 country-by-year observations, the gender gaps were positive but very small ($0.01 \le d < 0.10$), in 122 they were small ($0.10 \le d < 0.20$), in 35 they were medium-sized ($0.20 \le d < 0.30$), and in five they were rather pronounced ($0.30 \le d < 0.40$ in Kuwait in 2001, 2006, and 2011, and d = 0.39 and d = 0.43 in Saudi Arabia in 2011 and 2016).

< Table 2 about here >

< Figure 1 about here >

Main Results

In our main analyses, we first regressed the gender gap on to the continuous time (in years) variable and country dummies. By implication, our estimation of the effect of time was based on the longitudinal variation within countries. The results of this country-fixed effects model indicated that the international gender gap widened over time by β = 0.0006 (p < .027) per year. This linear annual increase is depicted as regression line in Figure 2. However, the trends in international gender gaps may not be linear. For this reason, we replaced the continuous time variable with dummies for each year and used 1970 as the reference category. The results for this analysis provided some evidence for a nonlinear relationship, because we observed a monotonically increasing trend up to the year 2001 and a monotonically decreasing trend thereafter The observed difference between the reference year 1970 and 2001 corresponds to an effect of β = 0.086 (see column 1 in Table 3 and points in Figure 2). However, while there were statistically significant differences between each year and the reference year 1970, the differences between the years 1991, 2001, 2006, 2011, and 2016 were rather small. For this reason, we wish to emphasize that our main finding was the change in size of the international gender gap between 1970 and 1991. The gender gap changes in the more recent years were small and conclusions should be drawn with caution.

< Table 3 about here >

< Figure 2 about here >

To contextualize the main findings, we decomposed the variance in the gender gaps across all country-by-year observations by means of an analysis of variance with time (categorical), country, and the interaction between them. This analysis showed that 72% of the total variance in the gender gap in country-by-year observations related to differences between countries, 7% was linked to international differences between time points, and 21% pertained to the interaction between countries and time. Since we found that the main source of variance was between countries, we would advise against overinterpreting the extent of the observed international trend. The results indicate that countries do indeed deviate from the international gender gap trend.

Robustness Checks

Several alternative models using more restricted country samples confirmed the robustness of the main findings. The first robustness test related to a more homogenous set of economically developed OECD member states. We estimated the country-fixed effects model with the categorical time variable for the subset of OECD countries (Table 3, column 2). The results from this analysis were remarkably similar to those of the main analyses using the full set of countries. Further robustness checks related to how often countries participated—at least three (column 3), four (column 4), five (column 5), or six times (column 6). Obviously, we had a dramatically reduced sample size of only 30 country-by-year observations when considering data from just the four countries that participated in all six years. Interestingly, despite the reduced sample sizes, the results were qualitatively the same for the more restricted samples.

Discussion

This study investigated long-term trends in international gender gaps in reading achievement at the end of primary schooling and generated the following findings. First, we found that girls generally scored higher than boys in reading comprehension in most countries and at most points in time. Second, we found that the size of the gender gap varied vastly across countries and time. The lowest gender gap of d = -0.08—therefore indicating a small reading advantage of boys over girls—was found in the Netherlands in 1970. The highest gender gap of d = 0.43 was found in Saudi Arabia in 2016, indicating a reading advantage of girls over boys of almost a half standard deviation. This is a large gender gap, especially when considering that the standard deviation in reading achievement pertains to 213 observations of a diverse set of countries between 1970 and 2016. Third—when modeling a general linear trend of the gender gap across time and controlling for the differential participation of countries in the years of observation—we found a small increase in the gender gap. Further analyses provide tentative evidence that the international gender gap showed a nonlinear rather than a linear trend. We found that the gender gap increased by an effect size of d = 0.09 between 1970 and 2001 and then slightly decreased until 2016 by an effect size of d = 0.03. Fourth, a decomposition of the variance in the gender gaps across countries and time indicated that differences in the gender gap were explained by time-stable between-country differences (72% of the variance) and country-specific trends (21% of the variance) rather than by a general international trend (7% of the variance). This finding helps to put the trend results in perspective.

Our findings are well aligned with previous research that also showed that countries differ in the magnitude of gender gaps (e.g., Thorndike 1973; Raudenbush et al. 1994; Mullis et al. 2017). Furthermore, previous studies using samples from many years ago also showed rather small gender gaps (e.g., Thorndike 1973; Hyde and Linn 1988; Lietz 2006a). By contrast, some studies using more recent samples found pronounced gaps (e.g., Chatterji 2006; Lietz 2006a; Lynn and Mikk 2009). However, since the more recent studies cover a more diverse set of countries than the older ones, these findings are difficult to interpret in terms of international long-term trends. This is a major contribution of the present study which accounted for methodological and sample differences over time. In line with our findings, previous short-term trend studies observed heterogeneous trends across countries, as well (e.g., Martin et al. 2003; Mullis et al. 2017; OECD 2019).

Explanations and Implications

Various theoretical arguments and perspectives can be utilized to explain our findings (cf. e.g., Connell 2002; Maccoby and Jacklin 1974; Halpern 2012). First, the result that girls performed better on the reading tests in almost all countries and at almost all points in time may have several explanations. One possible explanation for this tendency of female reading advantages is innate differences in underlying verbal cognitive abilities (cf. Maccoby and Jacklin 1974; Rosén 1998; Halpern 2012). However, internationally and temporally stable other reasons cannot be ruled out, either. In any case, from our point of view, the large variance between countries is the more interesting finding, which also suggests that more mechanisms than general cognitive ability differences must be at work. In the same vein, the changes over time suggest that achievement gender gaps are shaped be the context in which children grow up.

The literature considers various cultural, political, and school enrolment characteristics of countries as possible explanations for the large between-country differences in reading gender gaps (cf. UNESCO 2019; World Economic Forum 2019; Rosén et al. 2022). Interestingly, we found the most pronounced reading advantages for girls in the United Arab Emirates (Abu Dhabi), Kuwait, Oman, and Saudi Arabia (all measurement points $d > 0.20$). These are countries that are geographically and culturally relatively similar. Another group of countries that showed larger reading gender gaps were North-European countries (the Nordic and Baltic countries, United Kingdom and Ireland), which are again geographically and culturally similar to some extent. Future research could study potential mechanisms behind these patterns and investigate (dis-)similarities in these societies and education systems (cf. e.g., Guiso et al. 2008; McDaniel 2010; van Langen et al. 2006). Another perspective here might be to look at school enrolment rates. If there are gender gaps in primary school enrolment in some country-by-year observations, this should shape the respective gender gaps in academic achievement (cf. Steinmann and Rutkowski in print). The fact that the student sample sizes were not gender-balanced in all observations in the present study might be related to gender gaps in school enrolment; it is, however, no direct measure.

Similarly, cultural, political, and school enrolment factors can serve as potential explanations for the country-specific *trends* that we observed. We found relatively pronounced differences between measurement points in Chile, Cyprus, Finland, Hong Kong, Hungary, Iran, the Netherlands, and Portugal. Future case studies of these trends could discuss potential mechanisms behind these in the light of specific cultural, political, and enrolment-related changes. In the same vein, another interesting area for future case studies lies in the observed within-country differences between education systems (e.g., Ontario and Quebec in Canada, Abu Dhabi and Dubai in the United Arab Emirates). Such cases could allow to tentatively study effects of educational policies in contexts with very similar cultural implications of gender. Overall, we observed that the reading gender gaps remained quite stable over time in most countries, when linked longitudinally. This finding is not surprising when assuming that societal characteristics such as gender stereotypes change only slowly (cf. Halpern 2012; Maccoby 1998). In a similar vein, we only observed a small international trend that did not explain a lot of variance between country-by-year observations. We found a small increase of the gender gap between 1970 and 2001 and a slight decrease since then. Based on the literature and the present study, we could only speculate about reasons for this international trend.

However, this linking study is descriptive in nature and does not investigate reasons for the observed patterns statistically. It does however provide several interesting findings that can inform the theoretical debates about potential causes of differences between countries and over time and it can serve as a basis for future research that tries to explain the variation between the country-by-year observations. Evidence on specific cultural and political factors that shape gender gaps in academic achievement will then be able to inform policy-makers and educational stakeholders. But even without additional inferential evidence, this study provides some important implications for educational stakeholders. Parents, teachers, and educational policy-makers should recognize that boys and girls perform mostly similar on cognitive ability and academic achievement tests, especially in some countries, and that there is much more variation within the gender groups, than between them (e.g., Hyde 2005; Lindberg et al. 2010; Mullis et al. 2017; OECD 2019). Inflated assumptions about gender differences and stereotypical beliefs about the role of gender for education can have unintended effects on both boys and girls (e.g., Hyde 2005; Pahlke et al. 2014; UNESCO 2017). The fact that countries vary so much in achievement gender gaps provides strong evidence against the hypothesis that innate gender differences cause reading gender gaps in primary school. Furthermore, our findings challenge assumptions that boys would be increasingly falling behind in education in general in many countries (cf. OECD 2015, 2017).

In the case of reading achievement, our study rather suggests longstanding, stable gender gaps when considering methodological and sample differences over time.

Limitations and Outlook

Our analyses extend the present state of research by investigating actual long-term trends in a robust international design for the first time. We focused on students at the end of primary school who were investigated with comparable measures across countries and over time in the IEA's international reading assessments. We accounted for the fact that the studies investigated different country samples over time by estimating country-fixed effects. This approach allowed us to identify trends based on within-country variations. However, there is a possible criticism that these within-country estimations were sometimes based on only two observations. To counter this criticism, we further investigated sub-samples of countries that participated in multiple cycles. These robustness tests confirmed the main findings.

A possible point of criticism concerns the limited number of 213 country-by-year observations. This is, however, a general problem in country-level analyses. Our study tried to cope with the sample size problem by using all available measurement points of international large-scale assessments on reading achievement at the end of primary school and all countries that participated in at least two cycles. However, at the same time, we believe that using data from these representative international studies has distinctive advantages over using smaller-scale regional data, as gender gaps are expected to vary from context to context (cf. Wagemaker et al. 1996; Connell 2002; Cooray and Potrafke 2011). As Hedges and Nowell (1995) put it: "Most work on sex differences and talent has relied on data collected from samples that were not representative of the nation as a whole. Reviews and meta-analyses of data from nonrepresentative samples are not necessarily any more representative than the studies on which they are based" (ibid, 41). Nevertheless, our findings base on a limited number of countries, especially in the earlier assessments, which limits the statistical power. Furthermore, high-income countries are overrepresented in the IEA studies, which should be considered in the interpretation of the results.

The present study established a common reading comprehension scale across countries and time. The advantage of this approach is that it allows researchers to estimate trends on the country level. On the other hand, this limits the scope of the present work to the reading comprehension scale that was constructed. As has been argued by other researchers, gender gaps in reading might differ by text type (e.g., document texts versus literary texts) or specific item formats (e.g. multiple choice versus constructed response) (Hyde and Linn 1988; Wagemaker et al. 1996; Rosén 2001). Furthermore, we focused on gender gaps in mean achievement and not on gender gaps in the variability of reading achievement as an outcome (cf. Rosén 1998; Machin and Pekkarinen 2008; Gray et al. 2019) or other completely different achievement domains such as mathematics (cf. Lindberg et al. 2010; Meinck and Brese 2019; Mejía-Rodríguez et al. 2020).

A possible methodological issue concerns the linking of the reading tests over time. We argued that, in order to compare gender gaps over time, achievement must be measured with a comparable metric at every time point. To achieve this, the present study built on a linking study, which used a concurrent calibration to put all achievement measures onto the same IRT scale (Strietholt and Rosén 2016). This approach used item parameters that maximized the fit across all country-by-year observations to achieve comparable achievement scores. A natural limitation of this approach is, however, that some of the bridges between the tests were only given in the Swedish extensions of the international design. Consequently, we need to assume that the relevant item parameters from Sweden do not differ systematically from the (unobserved) item parameters in other countries. While it is impossible to test this empirically, we believe it is a tenable assumption because items in international studies typically do not show much item-by-country interaction. For example, not a single item in PIRLS 2016 showed severe item-by-country interaction (Foy et al. 2017). Another methodological limitation concerns the linking error, which can be large when only a small sample of items from previous studies is integrated into the new studies (e.g., Robitzsch and Lüdtke 2018; Weeks et al. 2013). In the present study, this concerns particularly the link between RCS and RLS, although even here 21 items overlap (see Appendix A2). We therefore assume that our linking error is moderate (Strietholt and Rosén 2016).

Conclusions

One main conclusion of this study is that, in many countries and at many points in time, the gaps between boys' and girls' reading comprehension scores were quite small. This is an important finding considering that there are "serious costs of overinflated claims of gender differences […] in many areas, including work, parenting, and relationships" (Hyde 2005, 589). At the same time, a key finding is the large variation in the gender gaps between countries. This pronounced international variation can only be explained by between-country differences, for instance, regarding educational systems (e.g., Marks 2008; McDaniel 2010; Hermann and Kopasz 2019), school enrolment rates (Steinmann and Rutkowski in print), cultural values (e.g., Connell 2002; Guiso et al. 2008; Cooray and Potrafke 2011), or gender stereotypes (e.g., Nosek et al. 2009; Reilly 2012).

Our finding that the international gender gap was rather stable over 46 years is consistent with the notion that such cultural and societal characteristics can be expected to change slowly. We did, however, find some indications for between-country differences in gender gap trends, which could reflect differential developments within countries. For instance, changes in gendered reading behaviors might correlate with changes in achievement trends. It is, however, beyond the scope of the present paper to investigate the actual causes of gender gap differences between countries and over time or to derive recommendations to adjust reading interventions accordingly (e.g., Guiso et al. 2008; Nosek et al. 2009; Hermann and Kopasz 2019). We believe that the described findings on international long-term gender gap trends can, however, serve as a valuable starting point for future studies that seek to explain between-country differences in gender gaps (cf. overview by Rosén et al. 2022). The present study has thus made an important contribution to the state of research by thoroughly describing and decomposing gender gaps in reading comprehension at the end of primary school across all available countries since the beginning of international large-scale studies.

References

Archambault, Isabelle, Jacquelynne S. Eccles, and Mina N. Vida. 2010. “Ability Self-Concepts and Subjective Value in Literacy: Joint Trajectories from Grades 1 through 12.” *Journal of Educational Psychology* 102, no. 4: 804–816.

Assié-Lumumba, N’Dri, and Margaret Sutton. 2004. “Global Trends in Comparative Research on Gender and Education.” *Comparative Education Review* 48, no. 4: 345–352.

Binkley, Marilyn, Keith Rust, and Marianne Winglee, eds. 1994. *Methodological Issues in Comparative Educational Studies: The Case of the IEA Reading Literacy Study*. Washington, DC: National Center for Educational Statistics.

Borgonovi, Francesca, Artur Pokropek, François Keslair, Britta Gauly, and Marco Paccagnella. 2017. “Youth in Transition: How Do Some of the Cohorts Participating in PISA Fare in PIAAC?” *OECD Education Working Papers* 155: 1–117.

Chatterji, Madhabi. 2006. “Reading Achievement Gaps, Correlates, and Moderators of Early Reading Achievement: Evidence from the Early Childhood Longitudinal Study (ECLS) Kindergarten to First Grade Sample.” *Journal of Educational Psychology* 98, no. 3: 489–507.

Connell, Raewyn W. 2002. *Gender*. Oxford, UK: Polity Press.

Cooray, Arusha, and Niklas Potrafke. 2011. “Gender Inequality in Education: Political Institutions or Culture and Religion?” *European Journal of Political Economy* 27, no. 2: 268–280.

Elley, Warwick B. 1992. *How in the World do Students Read? IEA Study of Reading Literacy*. Hamburg: IEA.

Else-Quest, Nicole M., Janet S. Hyde, and Marcia C. Linn. 2010. “Cross-National Patterns of Gender Differences in Mathematics: A Meta-Analysis.” *Psychological Bulletin* 136, no. 1: 103–127.

Foy, Pierre, Michael O. Martin, Ina V. S. Mullis, and Liqun Yin. 2017. “Reviewing the PIRLS 2016 Achievement Item Statistics” In *Methods and Procedures in PIRLS 2016*, Edited by Michael O. Martin, Ina V. S. Mullis and Martin Hooper: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA).

Gray, Helen, Andrew Lyth, Catherine McKenna, Susan Stothard, Peter Tymms, and Lee Copping. 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, no. 1: 1–29.

Guiso, Luigi, Ferdinando Monte, Paola Sapienza, and Luigi Zingales. 2008. "Diversity. Culture, Gender, and Math." *Science* 320, no. 5880: 1164–1165.

Halpern, Diane F. 2012. *Sex Differences in Cognitive Abilities*. 4th ed. New York, NY: Psychology Press.

Hanushek, Eric A., Stephen Machin, and Ludger Woessmann, eds. 2011. *Handbook of the Economics of Education.* 3 vols. Handbook of the Economics of Education. Amsterdam: North Holland.

Hanushek, Eric A., and Ludger Woessmann. 2011. "The Economics of International Differences in Educational Achievement." In *Handbook of the Economics of Education.* vol. 3. Edited by Eric A. Hanushek, Stephen Machin and Ludger Woessmann. 3 vols., 89–200. Handbook of the Economics of Education. Amsterdam: North Holland.

Hedges, Larry V., and Amy Nowell. 1995. "Sex Differences in Mental Test Scores, Variability, and Numbers of High-Scoring Individuals." *Science* 269, no. 5220: 41–45.

Hermann, Zoltán, and Marianna Kopasz. 2019. "Educational Policies and the Gender Gap in Test Scores: A Cross-Country Analysis." *Research Papers in Education*: 1–22.

Huang, Chiungjung. 2013. "Gender Differences in Academic Self-Efficacy: A Meta-Analysis." *European Journal of Psychology of Education* 28, no. 1: 1–35.

Hyde, Janet S. 2005. "The Gender Similarities Hypothesis." *The American Psychologist* 60, no. 6: 581–592.

Hyde, Janet S., and Marcia C. Linn. 1988. "Gender Differences in Verbal Ability: A Meta-Analysis." *Psychological Bulletin* 104, no. 1: 53–69.

Jerrim, John, and Gemma Moss. 2019. "The Link Between Fiction and Teenagers' Reading Skills: International Evidence from the OECD PISA Study." *British Educational Research Journal* 45, no. 1: 181–200.

Kiefer, Thomas, Alexander Robitzsch, and Margaret Wu. 2016. "TAM: Test Analysis Modules. R Package Version 1.995-0.". https://CRAN.R-project.org/package=TAM.

Kim, Seock-Ho, and Allan S. Cohen. 2002. "A Comparison of Linking and Concurrent Calibration under the Graded Response Model." *Applied Psychological Measurement* 26, no. 1: 25–41.

Klecker, Beverly M. 2006. “The Gender Gap in NAEP Fourth-, Eighth-, and Twelfth-Grade Reading Across Years.” *Reading Improvement* 43, no. 1: 50–56.

Lavrijsen, Jeroen, and Ides Nicaise. 2015. “Social Inequalities in Early School Leaving: The Role of Educational Institutions and the Socioeconomic Context.” *European Education* 47, no. 4: 295–310.

Lietz, Petra. 2006a. “A Meta-Analysis of Gender Differences in Reading Achievement at the Secondary School Level.” *Studies in Educational Evaluation* 32, no. 4: 317–344.

———. 2006b. “Issues in the Change in Gender Differences in Reading Achievement in Cross-National Research Studies Since 1992: A Meta-Analytic View.” *International Education Journal* 7, no. 2: 127–149.

Lindberg, Sara M., Janet S. Hyde, Jennifer L. Petersen, and Marcia C. Linn. 2010. “New Trends in Gender and Mathematics Performance: A Meta-Analysis.” *Psychological Bulletin* 136, no. 6: 1123–1135.

Lynn, Richard, and Jaan Mikk. 2009. “Sex Differences in Reading Achievement.” *Trames. Journal of the Humanities and Social Sciences* 13, no. 1: 3–13.

Maccoby, Eleanor E. 1998. *The Two Sexes: Growing up Apart, Coming Together*. Harvard: Belknap Press.

Maccoby, Eleanor. E., and Carol N. Jacklin. 1974. *The Psychology of Sex Differences*. Stanford: Stanford University Press.

Machin, Stephen, and Tuomas Pekkarinen. 2008. “Assessment. Global Sex Differences in Test Score Variability.” *Science* 322, no. 5906: 1331–1332.

Marks, Gary N. 2008. “Accounting for the Gender Gaps in Student Performance in Reading and Mathematics: Evidence from 31 Countries.” *Oxford Review of Education* 34, no. 1: 89–109.

Martin, Michael O., Ina V. S. Mullis, Eugenio J. Gonzalez, and Ann M. Kennedy. 2003. *Trends in Children’s Reading Literacy Achievement 1991-2001: IEA’s Repeat in Nine Countries of the 1991 Reading Literacy Study*. Chestnut Hill, MA: Boston College.

Masters, Geoff N. 1982. “A Rasch Model for Partial Credit Scoring.” *Psychometrika* 47, no. 2: 149–174.

McDaniel, Anne. 2010. “Cross-National Gender Gaps in Educational Expectations: The Influence of National-Level Gender Ideology and Educational Systems.” *Comparative Education Review* 54, no. 1: 27–50.

McDonnell, Lorraine M. 1995. “Opportunity to Learn as a Research Concept and a Policy Instrument.” *Educational Evaluation and Policy Analysis* 17*, no.* 3: 305–322.

Meinck, Sabine, and Falk Brese. 2019. “Trends in Gender Gaps: Using 20 Years of Evidence from TIMSS.” *Large-scale Assessments in Education* 7, no. 1: 1–23.

Mejía-Rodríguez, Ana M., Hans Luyten, and Martina R. M. Meelissen. 2020. “Gender Differences in Mathematics Self-Concept Across the World: An Exploration of Student and Parent Data of TIMSS 2015.” *International Journal of Science and Mathematics Education*: 1–22.

Mullis, Ina V. S., Michael O. Martin, Pierre Foy, and Martin Hooper. 2017. *PIRLS 2016 International Results in Reading*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA).

Nosek, Brian A., Frederick L. Smyth, N. Sriram, Nicole M. Lindner, Thierry Devos, Alfonso Ayala, and Yoav Bar-Anan et al. 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, no. 26: 10593–10597.

OECD. 2009. *Equally Prepared for Life? How 15-Year-Old Boys and Girls Perform in School: Programme for International Student Assessment*. Paris: OECD Publishing.

———. 2015 *The ABC of Gender Equality in Education: Aptitude, Behaviour, Confidence*. Paris: OECD Publishing.

———. 2016. *PISA 2015 Results (Volume I): Excellence and Equity in Education*. Paris: OECD Publishing.

———. 2017. *The Pursuit of Gender Equality: An Uphill Battle*. Paris: OECD Publishing

———. 2019. *PISA 2018 Results (Volume II)*. Paris: OECD Publishing.

———. 2019. *Skills Matter: Additional Results from the Survey of Adult Skills.* OECD skills studies. Paris: OECD Publishing.

———. 2020. *PISA for Development: Results in Focus*. PISA in Focus, no. 91. Paris: OECD Publishing.

Pahlke, Erin, Janet S. Hyde, and Carlie M. Allison. 2014. “The Effects of Single-Sex Compared With Coeducational Schooling on Students’ Performance and Attitudes: A Meta-Analysis.” *Psychological Bulletin* 140, no. 4: 1042–1072.

Pansu, Pascal, Isabelle Régner, Sylvain Max, Pascale Colé, John B. Nezlek, and Pascal Huguet. 2016. "A Burden for the Boys: Evidence of Stereotype Threat in Boys' Reading Performance." *Journal of Experimental Social Psychology* 65: 26–30.

Raudenbush, Stephen W., Yuk F. Cheong, and Randall P. Fotiu. 1994. "Synthesizing Cross-National Classroom Effects Data: Alternative Models and Methods." In *Methodological issues in comparative educational studies: The case of the IEA Reading Literacy Study*. Edited by Marilyn Binkley, Keith Rust and Marianne Winglee, 234–273. Washington, DC: National Center for Educational Statistics.

Reilly, David. 2012 "Gender, Culture, and Sex-Typed Cognitive Abilities." *PloS One* 7, no. 7: e39904.

Retelsdorf, Jan, Katja Schwartz, and Frank Asbrock. 2015. ""Michael Can't Read!" Teachers' Gender Stereotypes and Boys' Reading Self-Concept." *Journal of Educational Psychology* 107, no. 1: 186–194.

Robinson, Daniel B., Jennifer Mitton, Greg Hadley, and Meagan Kettley. 2021. "Single-Sex Education in the 21st Century: A 20-Year Scoping Review of the Literature." *Teaching and Teacher Education* 106: 103462.

Robitzsch, Alexander, and Oliver Lüdtke. 2018. "Linking Errors in International Large-Scale Assessments: Calculation of Standard Errors for Trend Estimation." *Assessment in Education: Principles, Policy & Practice* 26, no. 4: 444-465.

Rosén, Monica. 1998. *Gender Differences in Patterns of Knowledge*. Vol. 124. Göteborg Studies in Educational Sciences. Göteborg: Acta Universitatis Gothoburgensis.

Rosén, Monica. 2001. "Gender Differences in Reading Performance on Documents across Countries." *Reading and Writing: An Interdisciplinary Journal* 14 (1–2): 1–38.

Rosén, Monica, Isa Steinmann, and Inga Wernersson. 2022. "Gender Differences in School Achievement." In *International Handbook of Comparative Large-Scale Studies in Education*, edited by Trude Nilsen, Agnes Stancel-Piątak, and Jan-Eric Gustafsson, 1–48. Springer International Handbooks of Education. Cham: Springer International Publishing.

Steinmann, Isa, and Leslie Rutkowski. in print. "The Link between Gender Gaps in School Enrollment and School Achievement." *Comparative Education Review*.

Stoet, Gijsbert, and David C. Geary. 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, no. 3: e57988.

Strietholt, Rolf, Monica Rosén, and Wilfried Bos. 2013. “A Correction Model for Differences in the Sample Compositions: The Degree of Comparability as a Function of Age and Schooling.” *Large-Scale Assessments in Education* 1 (1). https://doi.org/10.1186/2196-0739-1-1.

Strietholt, Rolf, and Monica Rosén. 2016. “Linking Large-Scale Reading Assessments.” *Measurement: Interdisciplinary Research and Perspectives* 14 (1): 1–26.

Tao, Hung-Lin, and Christos Michalopoulos. 2018. “Gender Equality and the Gender Gap in Mathematics.” *Journal of Biosocial Science* 50, no. 2: 227–243.

Thorndike, Robert L. 1973. *Reading Comprehension Education in Fifteen Countries: An Empirical Study*. International Studies in Evaluation 3. Stockholm: Almqvist & Wiksell.

TIMSS & PIRLS International Study Center. 2019. “Trends in Reading Achievement by Gender.”. http://timssandpirls.bc.edu/pirls2016/international-results/pirls/student-achievement/trends-in-reading-achievement-by-gender/.

UNDP. 2019. *Human Development Report 2019: Beyond Income, Beyond Averages, Beyond Today: Inequalities in Human Development in the 21st Century*. New York: Bernan.

UNESCO. 2016. *2016 GEM Gender Review: Creating Sustainable Futures for All.* Global Education Monitoring Report. Paris: UNESCO.

———. 2017. *Cracking the Code: Girls' Education in Science, Technology, Engineering and Mathematics (STEM).* Paris: UNESCO.

———. 2019. *Building Bridges for Gender Equality*. Paris: UNESCO.

Van Langen, Annemarie, Roel Bosker, and Hetty Dekkers. 2006. “Exploring Cross-national Differences in Gender Gaps in Education.” *Educational Research and Evaluation* 12, no. 2: 155–177.

Wagemaker, Hans, Karin Taube, Ingrid Munck, Georgia Kontogiannopoulou-Polydorides, and Michael O. Martin. 1996. *Are Girls Better Readers? Gender Differences in Reading Literacy in 32 Countries*. Amsterdam: International Association for the Evaluation of Educational Achievement.

Waxman, Geffrey. 2015. “Reading Scores and Gender.” *European Scientific Journal* 11, no. 11: 199–212.

Weeks, Jonathan P, Matthias von Davier, and Kentaro Yamamoto. 2013. "Design Considerations for the Program for International Student Assessment." In *Handbook of International Large-Scale Assessment. Background, Technical Issues, and Methods of*

Data Analysis, Edited by Leslie Rutkowski, Matthias von Davier and David Rutkowski, 259-75. Boca Raton: Chapman & Hall.

Wigfield, Allan, and Jacquelynne S. Eccles. 2000. “Expectancy-Value Theory of Achievement Motivation.” *Contemporary Educational Psychology* 25, no. 1: 68–81.

Wolter, Ilka, Edith Braun, and Bettina Hannover. 2015. “Reading Is for Girls!? The Negative Impact of Preschool Teachers’ Traditional Gender Role Attitudes on Boys’ Reading Related Motivation and Skills.” *Frontiers in Psychology* 6: 1–11.

World Economic Forum. 2019. *The Global Gender Gap Report 2020*. Geneva: World Economic Forum.

Tables

Table 1

Country and Student Sample Sizes in the Seven Source Studies

			Original sample	Samples in the present study	
			# countries	# countries	# students
Year	Acronym	Study	(1)	(2)	(3)
1970	RCS	Reading Comprehension Study	14	12	27,216
1991	RLS	Reading Literacy Study	27	23	77,919
2001	RLS & PIRLS	Reading Literacy Study & Progress in International Reading Literacy Study	37 (9 also in RLS)	34 (9 also in RLS)	138,267
2006	PIRLS	Progress in International Reading Literacy Study	45	43	167,388
2011	PIRLS	Progress in International Reading Literacy Study	55	51	228,087
2016	PIRLS	Progress in International Reading Literacy Study	56	50	282,095
all			*234*	*213*	*920,972*

Note. The present study only included countries that participated in at least two of the six years of observation.

Table 2

Descriptive Statistics of the Gender Gaps in the Six Years of Observation

		Gender gap distribution				
Year	# countries	Minimum	1st Quartile	Mean	3rd Quartile	Maximum
1970	12	-0.08	0.02	0.06	0.10	0.15
1991	23	0.07	0.10	0.14	0.16	0.25
2001	34	0.05	0.12	0.16	0.21	0.30
2006	43	0.06	0.11	0.15	0.20	0.33
2011	51	-0.02	0.11	0.15	0.18	0.39
2016	50	0.01	0.09	0.13	0.17	0.43
all	*62*	*-0.08*	*0.10*	*0.14*	*0.18*	*0.43*

Note. The gender gaps are reported in effect size *d*. Values above zero indicate a mean reading advantage of girls over boys.

Table 3

Model Results of Regressing Gender Gaps on Categorical Time Variable

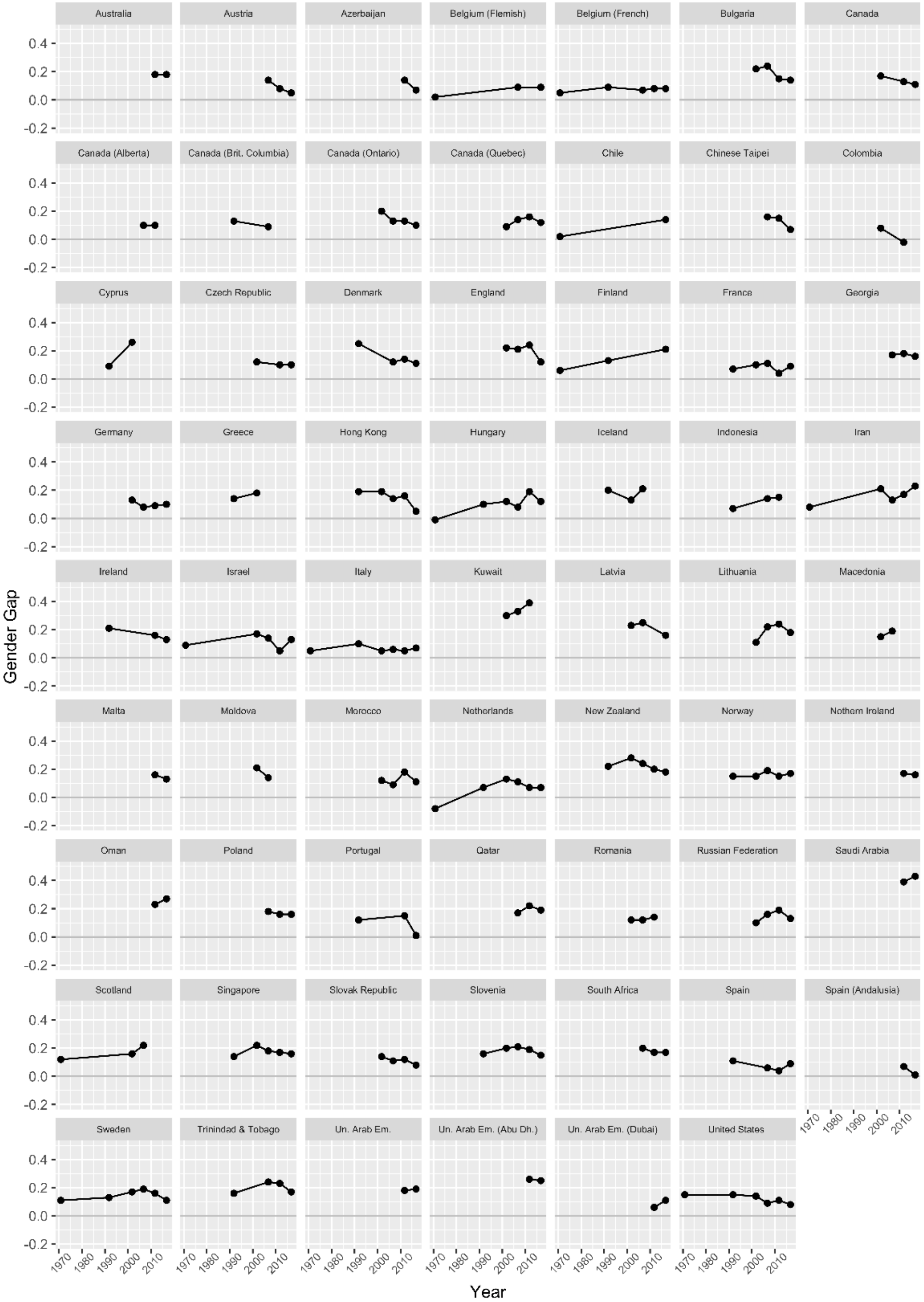
		Country subsamples				
	All countries	OECD countries	≥ 3 participations	≥ 4 participations	≥ 5 participations	6 participations
	(1)	(2)	(3)	(4)	(5)	(6)
Intercept (1970)	0.113**	0.071**	0.116**	0.068**	0.068**	0.066*
1991	0.070**	0.069**	0.070**	0.065**	0.059**	0.066*
2001	0.086**	0.075**	0.077**	0.076**	0.086**	0.078*
2006	0.077**	0.067**	0.074**	0.065**	0.064**	0.062*
2011	0.072**	0.046*	0.072**	0.063**	0.054**	0.072*
2016	0.052**	0.039	0.045**	0.041*	0.047**	0.046
Country-fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
n (country-by-year observations)	213	70	177	123	75	30

Note. Dependent variable is the gender gap in effect size *d*; the reference category is the year 1970; standard errors in parentheses; ** 1% significance level; * 5% significance level. The countries could participate in up to six measurement points between 1970 and 2016.

Figures

Figure 1

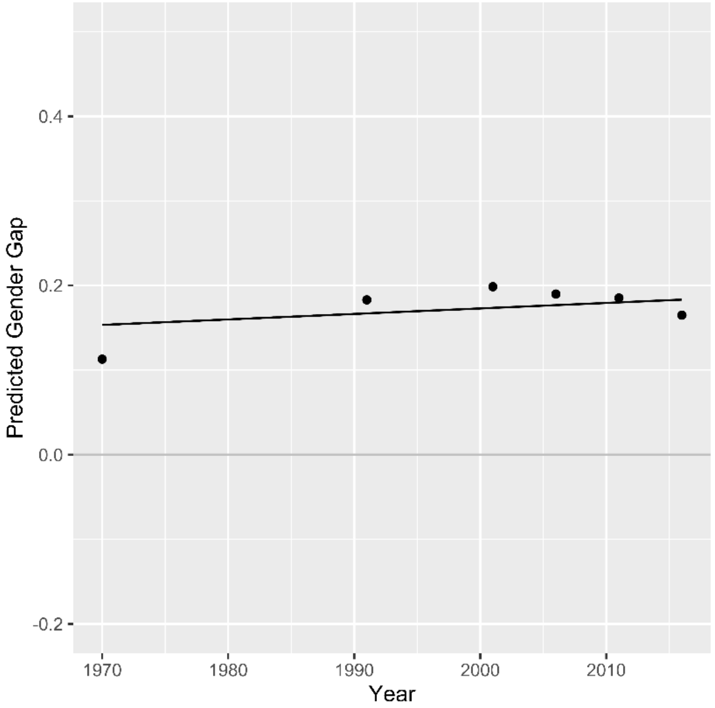
Descriptive Gender Gaps per Country and Year of Observation



Note. The gender gaps are in effect size *d*. Values above zero (depicted as grey horizontal lines) indicate a mean reading advantage of girls over boys.

Figure 2

Results of the Linear (Regression Line) and Nonlinear (Points) Reading Gender Gap Trend Estimations



Note. The gender gaps are in effect size *d*. Values above zero (depicted as horizontal line) indicate a mean reading advantage of girls over boys.

Appendix

Appendix A1

Countries' Study Participations with Grade Samples and Student Sample Sizes

Country	1970 RCS	1991 RLS	2001 RLS & PIRLS	2006 PIRLS	2011 PIRLS	2016 PRILS
Un. Arab Em. (Abu Dh.)					*n* = 3,300 Grade 4	*n* = 4,188 Grade 4
Un. Arab Em. (Dubai)					*n* = 4,818 Grade 4	*n* = 7,859 Grade 4
Un. Arab Em.					*n* = 11,634 Grade 4	*n* = 16,471 Grade 4
Australia					*n* = 4,852 Grade 4	*n* = 6,341 Grade 4
Austria				*n* = 4,056 Grade 4	*n* = 3,699 Grade 4	*n* = 4,360 Grade 4
Azerbaijan					*n* = 3,866 Grade 4	*n* = 5,994 Grade 4
Belgium (Flemish)	*n* = 717 Grade 4–6			*n* = 3,615 Grade 4		*n* = 5,198 Grade 4
Belgium (French)	*n* = 762 Grade 3–6	*n* = 2,695 Grade 4		*n* = 3,643 Grade 4	*n* = 2,961 Grade 4	*n* = 4,623 Grade 4
Bulgaria[1]			*n* = 2,580 Grade 4	*n* = 3,101 Grade 4	*n* = 4,197 Grade 4	*n* = 4,281 Grade 4
Canada (Alberta)				*n* = 3,381 Grade 4	*n* = 3,001 Grade 4	
Canada[1]			*n* = 6,142 Grade 4		*n* = 18,401 Grade 4	*n* = 18,245 Grade 4
Canada (Brit. Columbia)		*n* = 2,642 Grade 3		*n* = 3,329 Grade 4		
Chile	*n* = 1,461 Grade 3–6					*n* = 4,294 Grade 4
Colombia[1]			*n* = 3,846 Grade 4		*n* = 3,154 Grade 4	
Canada (Ontario)[1]			*n* = 3,205 Grade 4	*n* = 3,204 Grade 4	*n* = 3,614 Grade 4	*n* = 4,270 Grade 4
Canada (Quebec)[1]			*n* = 2,937 Grade 4	*n* = 2,981 Grade 4	*n* = 3,382 Grade 4	*n* = 3,179 Grade 4
Cyprus[1]		*n* = 1,494 Grade 4	*n* = 2,252 Grade 4			
Czechia[1]			*n* = 2,274 Grade 4		*n* = 3,613 Grade 4	*n* = 5,537 Grade 4
Germany[1]			*n* = 5,730 Grade 4	*n* = 6,307 Grade 4	*n* = 3,187 Grade 4	*n* = 3,959 Grade 4
Denmark		*n* = 3,463 Grade 3		*n* = 3,227 Grade 4	*n* = 3,650 Grade 4	*n* = 3,508 Grade 4
Spain (Andalusia)					*n* = 3,439 Grade 4	*n* = 4,169 Grade 4
England[1]			*n* = 2,379 Grade 5	*n* = 3,224 Grade[3] 5	*n* = 3,134 Grade 5	*n* = 5,095 Grade[3] 5
Spain		*n* = 8,228 Grade 4		*n* = 3,299 Grade 4	*n* = 6,829 Grade 4	*n* = 14,595 Grade 4
Finland	*n* = 1,293 Grade 3–5	*n* = 1,552 Grade 3				*n* = 4,896 Grade 4

			2001			
	1970	1991	RLS &	2006	2011	2016
Country	RCS	RLS	PIRLS	PIRLS	PIRLS	PRILS
France[1]		*n* = 1,874 Grade 4	*n* = 2,649 Grade 4	*n* = 3,524 Grade 4	*n* = 3,531 Grade 4	*n* = 4,767 Grade 4
Georgia				*n* = 3,548 Grade 4	*n* = 3,805 Grade 4	*n* = 5,741 Grade 4
Greece[1]		*n* = 3,516 Grade 4	*n* = 2,970 Grade 4			
Hong Kong[1]		*n* = 3,312 Grade 4	*n* = 3,791 Grade 4	*n* = 3,763 Grade 4	*n* = 3,086 Grade 4	*n* = 3,349 Grade 4
Hungary[2]	*n* = 4,841 Grade 4–6	*n* = 3,009 Grade 3	*n* = 8,209 Grade 3–4	*n* = 3,269 Grade 4	*n* = 4,142 Grade 4	*n* = 4,623 Grade 4
Indonesia		*n* = 3,167 Grade 4		*n* = 3,829 Grade 4	*n* = 3,826 Grade 4	
Ireland		*n* = 2,711 Grade 4			*n* = 3,602 Grade 4	*n* = 4,607 Grade 4
Iran[1]	*n* = 1,582 Grade 3–6		*n* = 5,534 Grade 4	*n* = 4,330 Grade 4	*n* = 4,577 Grade 4	*n* = 4,385 Grade 4
Iceland[2]		*n* = 3,976 Grade 3	*n* = 5,120 Grade 4	*n* = 4,032 Grade 4–5		
Israel[1]	*n* = 1,869 Grade 3–6		*n* = 2,988 Grade 4	*n* = 3,210 Grade 4	*n* = 3,343 Grade 4	*n* = 4,041 Grade 4
Italy[2]	*n* = 4,465 Grade 3–5	*n* = 2,232 Grade 4	*n* = 4,218 Grade 4	*n* = 2,861 Grade 4	*n* = 3,322 Grade 4	*n* = 3,940 Grade 4
Kuwait[1]			*n* = 6,470 Grade 4	*n* = 3,201 Grade 4	*n* = 2,705 Grade 6	
Lithuania[1]			*n* = 1,865 Grade 4	*n* = 3,755 Grade 4	*n* = 3,696 Grade 4	*n* = 4,317 Grade 4
Latvia[1]			*n* = 2,272 Grade 4	*n* = 3,340 Grade 4		*n* = 4,157 Grade 4
Morocco[1]			*n* = 2,360 Grade 4	*n* = 2,617 Grade 4	*n* = 11,977 Grade 4–6	*n* = 5,489 Grade 4
Moldova[1]			*n* = 2,674 Grade 4	*n* = 3,252 Grade 4		
Macedonia[1]			*n* = 2,776 Grade 4	*n* = 3,203 Grade 4		
Malta					*n* = 2,849 Grade 5	*n* = 3,647 Grade 5
Northern Ireland					*n* = 2,848 Grade 4	*n* = 3,693 Grade 4
Netherlands[1]	*n* = 1,611 Grade 6	*n* = 1,700 Grade 3	*n* = 3,094 Grade 4	*n* = 3,360 Grade 4	*n* = 3,188 Grade 4	*n* = 4,206 Grade 4
Norway[1]		*n* = 2,444 Grade 3	*n* = 2,595 Grade 4	*n* = 4,553 Grade 4–5	*n* = 2,531 Grade 4	*n* = 8,586 Grade 4–5
New Zealand[2]		*n* = 3,016 Grade 5	*n* = 3,067 Gr.[3] 4.5–5.5	*n* = 5,034 Gr.[3] 4.5–5.5	*n* = 4,471 Gr.[3] 4.5–5.5	*n* = 5,646 Gr.[3] 4.5–5.5
Oman					*n* = 8,276 Grade 4	*n* = 9,234 Grade 4
Poland				*n* = 3,903 Grade 4	*n* = 3,966 Grade 4	*n* = 4,413 Grade 4
Portugal		*n* = 2,778 Grade 4			*n* = 3,247 Grade 4	*n* = 4,642 Grade 4
Qatar				*n* = 5,365 Grade 4	*n* = 3,275 Grade 4	*n* = 9,077 Grade 4
Romania[1]			*n* = 2,706 Grade 4	*n* = 3,428 Grade 4	*n* = 3,701 Grade 4	

Country	1970 RCS	1991 RLS	2001 RLS & PIRLS	2006 PIRLS	2011 PIRLS	2016 PRILS
Russian Federation[1]			*n* = 3,071 Grade 3–4	*n* = 3,804 Grade[3] 3–4	*n* = 3,549 Grade 4	*n* = 4,577 Grade 4
Saudi Arabia					*n* = 3,581 Grade 4	*n* = 4,741 Grade 4
Scotland[1]	*n* = 2,121 Grade 4–6		*n* = 2,058 Grade 5	*n* = 3,018 Grade 4		
Singapore[2]		*n* = 7,326 Grade 3	*n* = 8,856 Grade 3–4	*n* = 5,116 Grade 4	*n* = 5,018 Grade 4	*n* = 6,488 Grade 4
Slovak Republic[1]			*n* = 2,863 Grade 4	*n* = 4,304 Grade 4	*n* = 4,467 Grade 4	*n* = 5,451 Grade 4
Slovenia[2]		*n* = 3,298 Grade 3	*n* = 3,718 Grade 3	*n* = 4,289 Grade 3–4	*n* = 3,583 Grade 4	*n* = 4,499 Grade 4
Sweden[2]	*n* = 1,951 Grade 3–4	*n* = 4,301 Grade 3	*n* = 16,592 Grade 3–4	*n* = 3,536 Grade 4	*n* = 3,683 Grade 4	*n* = 4,525 Grade 4
Trinidad & Tobago		*n* = 3,683 Grade 4		*n* = 3,161 Grade[3] 5	*n* = 3,138 Grade[3] 5	*n* = 4,177 Grade[3] 5
Chinese Taipei				*n* = 3,675 Grade 4	*n* = 3,429 Grade 4	*n* = 4,326 Grade 4
America[2]	*n* = 5,418 Grade 3–6	*n* = 6,546 Grade 4	*n* = 4,663 Grade 4	*n* = 4,131 Grade 4	*n* = 10,142 Grade 4	*n* = 4,425 Grade 4
South Africa				*n* = 11,702 Grade 5	*n* = 2,819 Grade[3] 4	*n* = 5,282 Grade[3] 4

Note. [1] means that the country participated only in PIRLS and [2] means that the country participated in both RLS and PIRLS in 2001. [3] means that grade information was taken from international reports instead of raw datasets. We only included countries that participated in at least two of the six assessment cycles. Empty cells imply that a country did not participate in the respective cycle.

Appendix A2

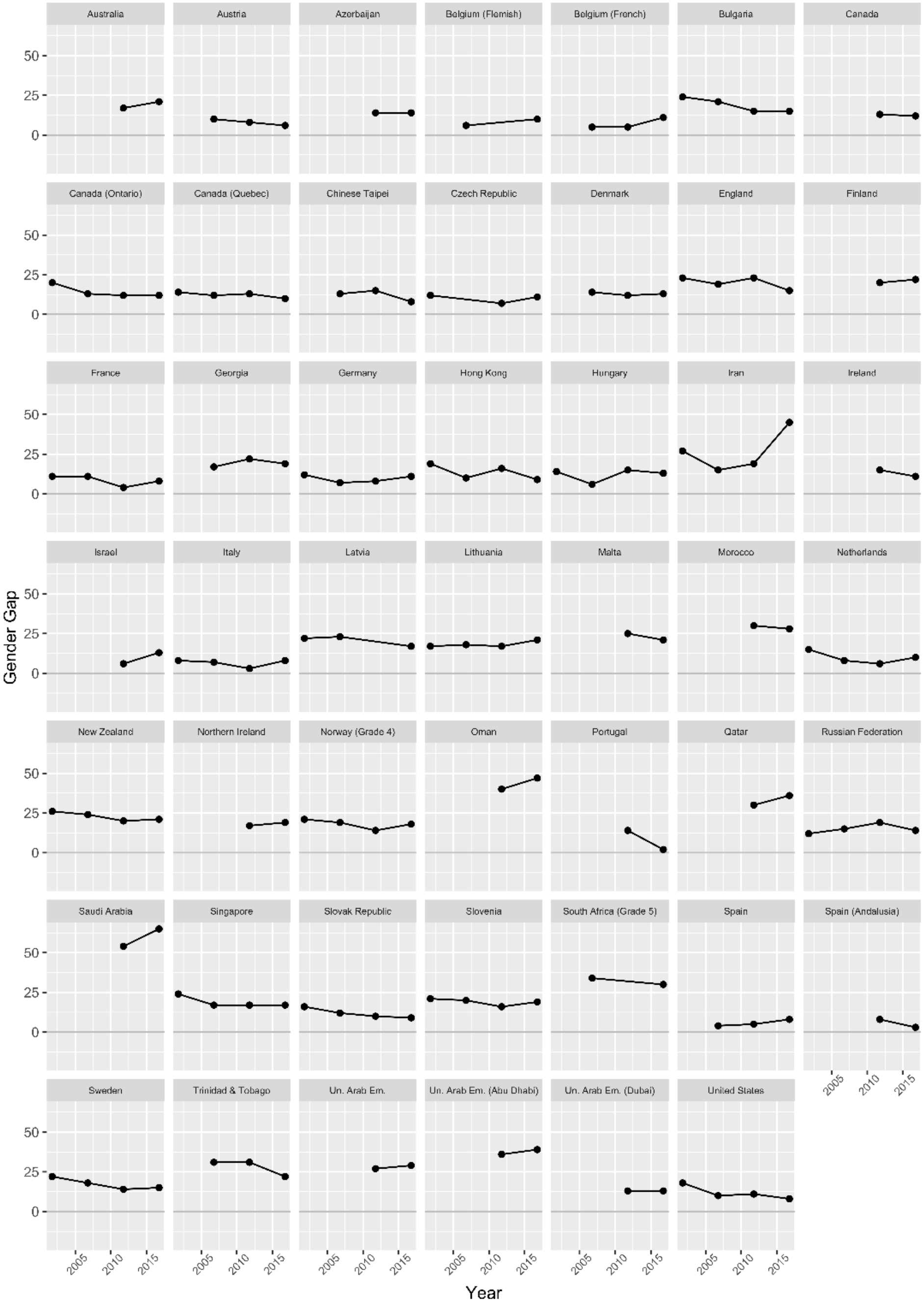
Text Passages and Number of Items over Time

		1970	1991	2001		2006	2011	2016
Text passage	Text type	RCS	RLS	RLS	PIRLS	PIRLS	PIRLS	PIRLS
Poet	narrative	ID						
Pole	narrative	ID						
Seal	expository	ID						
Ox	expository	ID						
Marmots	expository	ID						
Erneke	narrative	ID	SE	SE				
Tailor	expository	ID	SE	SE				
Plant	expository	ID	SE	SE				
Marmots	expository	ID	ID	ID	SE			
The Bird and the Elephant	narrative		ID	ID				
Grandpa	narrative		ID	ID	SE			
A Shark Makes Friends	narrative		ID	ID	SE			
No Dogs is not Enough	narrative		ID	ID	SE			
Postcard	expository		ID	ID	SE			
What is Quicksand?	expository		ID	ID	SE			
The Walrus	expository		ID	ID	SE			
How to Read the Age of a Tree	expository		ID	ID	SE			
The Upside-Down Mice	narrative				ID			
River Trail	expository				ID			
The Little Lump of Clay	narrative				ID	ID		
Antarctica	expository				ID	ID		
Flowers on the Roof	narrative				ID	ID	ID	ID
Leonardo da Vinci	expository				ID	ID	ID	ID
Fly Eagle	narrative					ID	ID	
Day Hiking	expository					ID	ID	
Shiny Straw	narrative					ID	ID	ID
Sharks	expository					ID	ID	ID
The Empty Pot	narrative						ID	ID
Where's the Honey?	expository						ID	ID
Oliver and the Griffin	narrative							ID
Pemba Sherpa	narrative							ID
Icelandic Horses	expository							ID
How Did We Learn to Fly?	expository							ID

Note. ID means international design (i.e., text passage and items assessed in all countries). SE means Swedish extension (i.e., text passages and items only assessed in Sweden). One item of the Marmots text passage was only used in 1970 and not later on.

Appendix A3

Gender Gap Trends in PIRLS 2001, 2006, 2011, and 2016 in Raw Score Points as Reported in Exhibit 1.6 in Mullis et al. (2017)



Note. The gender gaps are reported in mean score differences. The PIRLS reading scale has an international mean of 500 and a standard deviation of 100. Values above zero (depicted as horizontal lines) indicate a mean reading advantage of girls over boys.