Lorenz, Theresa; Schneebaum, Alyssa

Article

Does early educational tracking contribute to gender gaps in test achievement? A cross-country assessment

Journal of Economics and Statistics

Provided in Cooperation with:
De Gruyter Brill

Suggested Citation: Lorenz, Theresa; Schneebaum, Alyssa (2024) : Does early educational tracking contribute to gender gaps in test achievement? A cross-country assessment, Journal of Economics and Statistics, ISSN 2366-049X, De Gruyter Oldenbourg, Berlin, Vol. 244, Iss. 1/2, pp. 5-36, https://doi.org/10.1515/jbnst-2022-0005

This Version is available at:
https://hdl.handle.net/10419/333273

Theresa Lorenz and Alyssa Schneebaum*

Does Early Educational Tracking Contribute to Gender Gaps in Test Achievement? A Cross-Country Assessment

https://doi.org/10.1515/jbnst-2022-0005
Received February 7, 2022; accepted March 27, 2023

Abstract: On average, boys score higher on math achievement tests and girls score higher in reading; these gaps increase between primary and secondary school. Using PISA, PIRLS, and TIMSS data, we investigate the role of early educational tracking (sorting students into different types of secondary schools at an early age) on gender gaps in test achievement in a cross-country difference-in-differences framework. We find strong evidence that early tracking increases gender differences in reading. For math test scores, we do not find consistent evidence that early tracking contributes to the gender gap.

Keywords: PISA, TIMSS, PIRLS, gender gaps, educational systems, early tracking

JEL Classification: I24, I28, J16, H52

1 Introduction

Results from large-scale international achievement tests such as PISA (Programme for International Student Assessment), TIMSS (Trends in International Mathematics and Science Study), and PIRLS (Progress in International Reading Literacy Study) reveal gender gaps in student test scores. On average, girls score higher than boys in reading, while boys outperform girls in math (Guiso et al. 2008; OECD 2009a; OECD 2016b; Mullis et al. 2015). At the mean, the reading gap is more pronounced than the mathematics

The views and results presented in this paper are those of the authors and do not necessarily represent the official opinions of the Austrian Central Bank or the Eurosystem.

Article Note: This article is part of the special issue "Gender Economics" published in the Journal of Economics and Statistics. Access to further articles of this special issue can be obtained at www.degruyter.com/journals/jbnst.

***Corresponding author: Alyssa Schneebaum**, Department of Economics, Vienna University of Economics and Business, Welthandelsplatz 1, 1020 Vienna, Austria, E-mail: alyssa.schneebaum@wu.ac.at
Theresa Lorenz, Financial Literacy Division, Austrian Central Bank (OeNB), Vienna, Austria

gap, but high-scoring boys perform better than high-scoring girls in math (Baye and Monseur 2016; Bedard and Cho 2010). A particularly worrying pattern emerges from the data: the gaps in test scores that exist at the end of primary school become more pronounced as school children age. The gap at the end of lower secondary school is larger than the gap at the end of primary school (Baye and Monseur 2016; Fryer et al. 2010). Cultural differences across countries may help explain the widening gaps. For example, the gap in math is less pronounced in gender-egalitarian countries (Else-Quest et al. 2010; Stoet and Geary 2015). Educational institutions across countries may also differently affect girls' and boys' schooling performance. The level of standardization in school (Connor et al. 2013), the sex composition in classes (Lee and Bryk 1986), and variation in students' socio-economic background (Legewie and DiPrete 2012) have all been found to impact girls' and boys' performance differently.

In this paper we focus our analysis on the effect of one particular institutional school setting: the age at which students are first sorted into different types of secondary school. In early tracking countries, lower secondary schools sort students into tracks depending on their primary school performance; in late tracking countries, all students attend the same lower secondary school with similar educational trajectories.[1] Several studies have investigated the impact of early tracking on overall educational inequality, finding that it disadvantages the lowest-performing students (Ammermüller 2005; Hanushek and Wößmann 2006; Montt 2011). Other assessments of early tracking show that it has negative effects on political engagement (Van de Werfhorst 2017); leads to worse test performance for the children of immigrants who do not speak the language of the testing country at home (Ruhose and Guido 2016); and, in contrast to these other findings, has slightly equalizing effects on the health outcomes of people across educational classes (Delaruelle et al. 2019).

The contribution of the present paper is to study the effect of early tracking on the gender gap in math and reading test scores. Existing literature on the topic of early tracking and gender test score gaps has not reached a consensus finding. For reading, the findings in van Hek et al. (2019) suggest that early tracking gives an advantage to boys, while the analysis in Hermann and Kopasz (2019) suggests the opposite. For math, too, the results are mixed. Bedard and Cho (2010) find that early tracking is associated with a higher gender gap in achievement; Hermann and Kopasz (2019) find that early tracking corresponds with a lower gender gap; and Bodovski et al. (2020) find no statistically significant relationship between early

1 The age of first tracking usually doesn't differ across regions within a country. Notable exceptions include Germany, the effects of whose region-specific tracking systems have been studied widely (e.g. Schindler and Bittmann 2021; Traini et al. 2021). Other exceptions have given the context for studies looking at the relationship between early versus late tracking and gendered educational inequality in Finland (Pekkarinen 2008) and Northern Ireland (Maurin and McNally 2007).

tracking and the gender gap in math scores. A main reason for the discrepancy in results is the inconsistency in data and especially method applied. None of these papers employ an empirical strategy that could lead to a causal interpretation of the results. Their insights are very important to the literature, but when assessing the *effect* of early tracking on gender gaps in test performance, several other steps need to be taken. In particular, the few existing studies looking at this question either leave out fundamental variables that can help explain achievement gaps (such as per-pupil expenditures or class time spent working on a particular subject); do not control for existing gaps in test scores at the end of primary school; or do not include the information provided by the complex survey design in the data.

In this paper, we use a cross-country difference-in-differences approach, using variation in countries' age of first tracking as well as information on test scores from before tracking (primary school, using TIMSS and PIRLS data) and after tracking (secondary school, using PISA data) to see the effect of early tracking. empirical approach is very similar to that taken by Ruhose and Guido (2016), who study the effect of early tracking on the gap in test scores for children of immigrants versus native-born parents.

There are two theories to help frame our understanding of why early tracking could contribute to a gender gap in test scores. first is the "maturity hypothesis," which can explain girls' higher scores in reading. There are two streams to this hypothesis, both based on the large literature showing that boys mature later than girls (e.g. DiPrete and Jennings 2012; Tanner 1990). The streams differ in their assumptions about the age at which the maturity gap emerges. The first stream says that the maturity gap emerges at a young age, before even early tracking occurs, giving girls in early tracking countries better chances of being selected for a more rigorous academic track. Since all students perform better when they are grouped with higher achievers (Altermatt and Pomerantz 2005; Hanushek and Kain 2003; Huang 2009; Robertson and James 2003), girls placed onto higher tracks because of their greater maturity will benefit from early tracking more than less-mature boys (Bedard and Cho 2010; Jürges and Schneider 2011). Early sorting seems indeed to be related to a large gap in reading scores: even controlling for test achievement at the end of primary school, the gap in reading in early tracking countries is larger than in late tracking countries (Jürges and Schneider 2011). The second stream of the maturity hypothesis says that the maturity gap emerges later.[2] In this case, it will be *late* tracking instead that benefits girls (Anderson et al. 2001; Keulers et al. 2010; Pekkarinen 2008). Depending on when the maturity gap emerges and its alignment with a country's tracking age, girls can have a larger or smaller advantage from

2 Adolescent boys have lower grades and weaker educational aspirations, suggesting that the maturity gap may indeed be most pronounced in these later years (Dubas et al. 1991; Daniel et al. 1982).

tracking age. The streams of the maturity gap hypothesis help us understand the size and timing of girls' better performance in reading, but cannot explain why boys score higher in mathematics.

A second theory can help explain girls' lower average performance in mathematics: that of gender roles and socialization that prescribe math as a "boy" subject. Girls are less likely to participate in mathematics courses and boys are underrepresented in non-science subjects in secondary school (Langen et al. 2008; Pinxten et al. 2012), in part because students adapt their focus in schooling based on gender roles prescribed by parental, teacher, and peer beliefs about gender identity (Gallagher and Kaufman 2004; Hadjar et al. 2014; Ma 2001). We suggest that socialization can help explain why models of major choice based solely on student ability predict that more girls should be in science, mathematics, economics, and engineering majors while more boys should be in humanities, psychology, and social science-orientated tracks than is actually the case (Hyde and Mertz 2009; Turner and Bowen 1999). Indeed ethnographers observe that high school students classify themselves into social categories and try to find the best match between their own and their peers' social identity (Eckert 1989). Moreover, teachers' prejudices regarding girls' and boys' abilities in certain areas might affect how they treat and grade boys and girls (Hörstermann et al. 2010; Li 1999; Ziegler et al. 1998). Early tracking can lead to disproportionate sorting into gender-typical fields because young children may be more inclined to follow the socially assigned path for their biological sex, which would increase gender gaps in both math (in favor of boys) and reading (in favor of girls).

2 Empirical Strategy

To study the link between early tracking and gender gaps in test scores, we use a difference-in-differences (DiD) approach, following the progression in the relevant literature from Hanushek and Wößmann (2006), Ruhose and Guido (2016), and Hermann and Kopasz (2019). Hermann and Kopasz (ibid.) are alone in having applied cross-country DiD estimation methods to the analysis of early tracking and gender differences in test scores. They used multilevel modelling and found that early tracking raises girls' achievement test scores. The present study goes beyond the work of Hermann and Kopasz (ibid.) in four ways. First, we employ the appropriate weights and multiple imputation scheme on the data in our analysis, which is necessary to maximize the accuracy of the information in the data sets.[3] Second, we

3 Hermann and Kopasz (2019) do not mention the use of survey weights and multiple imputation in their study, so we assume that they did not use them. Most studies on early tracking and gender inequality apply weights and clustered standard errors, but only van Hek et al. (2019) correctly

use more recent waves of the data (2015 for PISA and 2011 for TIMSS/PIRLS); in the data section below we discuss why we choose to use these waves instead of the currently most recent available data, TIMSS 2019, PIRLS 2016, and PISA 2018. Third, we include a battery of important robustness checks to verify and deepen the understanding of our results. Fourth, we include important country- and student-level information that justify the use of DiD to identify causal effects.

Follow the same empirical strategy to test for the effect of early tracking on test scores gaps as Ruhose and Guido (2016). We define countries as early tracking if they track students for the first time before the age of 15 (we test the sensitivity of this cut-off point in robustness checks below). We estimate the following equation:

$$\begin{aligned} Y_{ics} = \; & \beta_0 + \beta_1 \times ET_c \times SEC_s \times G_i + \beta_2 \times ET_c \times G_i + \beta_3 \times ET_c \times SEC_s + \\ & \beta_4 \times G_i \times SEC_s + \beta_5 \times G_i + \beta_6 \times SEC_s + \\ & \mathbf{X}'\sigma + \Sigma\,(\mathbf{X} \times SEC_s) + \mathbf{Z}'\pi + \Pi\,(\mathbf{Z} \times G_i) + \mu_c + v_{ics} \end{aligned} \tag{1}$$

where Y_{ics} is the test score of individual i in country c; the index s represents the test score in secondary school. ET_c is an indicator equal to 1 if the country is an early tracking country. SEC_s is an indicator equal to 1 if the individual is in secondary school; this variable can be interpreted as a grade fixed effect. G_i is an indicator equal to 1 for the lower achieving sex, that is, it is equal to 1 for girls for the models predicting math scores and equal to 1 for boys in the models predicting reading scores. μ_c is a country fixed effect and captures all country-specific time invariant factors. $\mathbf{X}$ is a vector that includes individual-level background variables, capturing information on student age; minutes spent in mathematics and reading in school; the number of books at home; and father's and mother's education. interact all of these individual-level background variables with the indicator of whether a student is in secondary school via $SEC_s \times X$.

textbfZ is a vector of four country-level time-variant control variables: GDP growth; the year-on-year change in educational expenditures; the share of female teachers; and an index measure of a country's gender equality. We interact these country-level measures with the gender indicator to test for different effects of these country-level characteristics on male versus female students via $\mathbf{Z} \times G_i$.

For causal identification of β_1, one has to make the following three identifying assumptions. First, there are no unobservable factors correlating (a) the existence of early tracking and (b) gender differences in primary school test achievement at different points in the school year. The "different points in the school year" is relevant because both TIMSS and PIRLS exams are conducted during but not exactly at the end of the school year. The assumption implies that there should be no structural

incorporate the plausible values (that is, the plausible multiple imputation method) also recommended by the OECD (2009b), though not in a causal DiD setting.

differences between early and late tracking countries that change the gender gap *depending on when during the school year students are tested.* This first assumption would only be violated if, for example, parents in early tracking countries push girls and boys differently at the end of primary school, which seems unlikely.

Second, there can be no unobservable factors that differently affect the evolution of boys' and girls' test scores in early versus late tracking countries that are not already present at the end of primary school. There are several potential examples of such issues, such as if per-pupil expenditures between primary and secondary school in early tracking countries increased or decreased to a different extent than in late tracking countries and one gender benefited more from this change in expenditures. Other potential confounding time effects might be if the share of female teachers changed at different rates in early versus late tracking countries and the resulting effects on students differ by gender (Diefenbach and Klein 2002), or if GDP grew differently in late versus early tracking countries and these changes in GDP growth between grade four and grade eight affected the performance of girls and boys differently. These potential scenarios are the impetus for why we include controls for GDP growth, education expenditure change, female teacher share, and overall gender inequality in a country in our empirical models. Inclusion of these variables eliminates time-effect distortions, and is one of the most important methodological improvements over earlier studies. Only Bodovski et al. (2020) studying the relationship between school differentiation and the gender gap in math and science (but not reading) include similar variables (though in the context of a non-causal analysis, not taking differences at the end of primary school into account).[4]

Third and finally, we need to assume that there are no structural differences between early and late tracking countries that could affect the gender gap depending on whether students at the age of 15 are still in lower secondary school. Even though this assumption is widely ignored in the literature, it is crucial when analyzing the determinants of gender gaps in test achievement.[5] We therefore include information

4 A potential related issue is migration. We would have confounding time effects if there are differences in the immigration rates between early and late tracking countries that are not already captured in β_2 and if these differences affect boys and girls differently. This is unlikely, but in PIRLS 2006 and TIMSS 2007 contain information on parents' country of birth and we use it in robustness checks below.

5 An example will illustrate: in Austria (an early tracking country), students usually choose between different tracks a second time at the age of 14. At the age of 15, when PISA is conducted, some students are already on tracks that are highly specialized in certain fields, such as technical or language-based tracks with completely different curricula. The tracks are highly gender-differentiated: technical schools comprise almost 75 % males, while social tracks comprise almost 85 % females (Austria 2017). In many late tracking countries, however, the age of first selection is first at 16 (as shown in Table 1), which means that the schools have a lower degree of gender differentiation at the age of 15, when PISA is conducted. Early tracking might thus be correlated with gender differentiation at the age of 15.

on "minutes spent in mathematics" and "minutes spent in test language" to the models. As such, we compare students with the same time spent on these subjects; the variables serve as a proxy for the degree of gender segregation in mathematics and reading at the age of 15 and arguably fulfills assumption three.

School systems across countries differ in important ways, and in ways that may be related to gender gaps in achievement scores in early versus late tracking countries. By controlling for cross-country differences in GDP growth, changes in school expenditures, the share of female teachers, and an index measure of gender equality, as well as student-specific number of minutes spent in math and reading, we arguably deal with most factors that would question the validity of a causal interpretation of our results. In the next section we describe our data in more detail before turning to the results in Section 4.

3 Data

We use the PIRLS 2011, TIMSS 2011, and PISA 2015 datasets for the main analysis. PISA data come from the OECD while PIRLS (reading) and TIMSS (math) come from the International Association for the Evaluation of Educational Achievement (IEA). All three studies were first conducted in the late 1990s or early 2000s and were subsequently repeated every few years (PISA every three years, PIRLS every four years, and TIMSS every five years) (IEA 2019a; IEA 2019b; OECD 2019).

Data from previous waves are used in some robustness checks. Data from newer waves are available but are not appropriate for our empirical strategy. Our main models compare PIRLS and TIMSS data on primary school students in 2011 with PISA data on secondary school students in 2015. The timing of these surveys implies that the students we compare were born in the same year. Therefore, comparing the students in these datasets allows us to eliminate any differences in students related to birth cohort, such as social norms about gender while the child was very young. The timing of newer waves of the data do not offer the opportunity to control for birth cohort effects. The newest data available are on primary school students in 2016 (PIRLS) and 2019 (TIMSS) and secondary school students in 2018 (PISA). These data do not line up to control for birth cohort effects. The other main potential source of bias is time effects: much changed between 2011 and 2015 (e.g. the economy in most sampled countries was stronger in 2015 than in 2011) that may have affected school performance of boys and girls differently. We deal with this by controlling for time-variant factors, such as per-pupil expenditures and GDP

Without controlling for this fact, β_1 might capture the effect of gender segregation at the age of 15 rather than the effect of early tracking.

growth, and interacting these with gender. Newer waves of the data would also have time-specific effects (especially comparing PIRLS 2016 to PISA 2018), while also suffering from birth cohort effects. We thus use older data because they allow us to at least eliminate birth cohort effects.

The most important difference between the PISA and the IEA datasets is that the target population in PISA is 15-year-old students, whereas the TIMSS and PIRLS samples comprise grade four students, who are 10 or 11. All datasets contain a full set of responses from individual students, school principals, and parents to very similar questions. Additionally, TIMSS (which asks about math, but not reading) is the only of the three surveys conducted at grades four and eight. We exploit this fact to perform the analysis within two available grades in TIMSS in a robustness check below. For all analyses, we standardize the data such that the mean scores are 500 with a standard deviation of 100. We use senate weights throughout the analysis, which means that each country contributes equally to the results. This is important in our context because the policy intervention (early tracking) is done on the country level; for the overall analysis of the policy, the effects of the policy in each country should be counted the same. Otherwise the effects of the policy in larger countries, like Germany, would play a stronger role in the conclusions about the effect of the policy.

To limit the range of cultural differences across countries in the sample, the main analysis includes only European countries. Robustness checks based on additional OECD countries are shown below. Some European countries were dropped due to important missing country-level data or because they were not included in one of the three main datasets. If important individual-level data were missing, we impute them using country averages. The 20 countries yield more than 200,000 individual-level observations.

The multivariate regressions described in equation (1) include the following individual-level control variables. “Age” indicates student’s age at the time of the test. “Books” shows students’ response to the question “how many books do you have at home?”; the response categories were “none or few books,” “one bookshelf,” “one bookcase,” “two bookcases,” and “three or more bookcases.” We align these answers into categories 1–5 in our data. “Books” together with the parents’ education variables act as family background proxy variables. We control for dummy variables indicating the highest completed ISCED level of education of both the father and the mother, measured as primary, secondary, or tertiary education. In TIMSS and PIRLS, a variable on time spent in math and language is derived from teacher’s response to the question, “In a typical week, how much time do you spend on language/mathematics instruction and/or activities with the students?” In PISA, students answer the question, “How many class periods per week are you typically required to attend for the subjects in test language/mathematics?”

On the country level, we derive control variables from other sources. The OECD provides data on the age of first tracking (OECD 2013, Figure IV.2.4, p. 78; OECD 2016a, Figure II.5.8, p. 167). These data were double-checked with UNESCO's World Data on Education, which provide detailed information on different European education systems (UNESCO Institute for Statistics 2013). Following Hanushek and Wößmann (2006) and Ruhose and Guido (2016), we define a country as early tracking if the first tracking occurs before the age of 15. Countries that track students at 15 or later are categorized as late tracking countries. Based on this definition, there are 10 late tracking countries and 10 early tracking countries in the main model. Table 1 shows the age of first tracking for the baseline sample countries and additional countries used in robustness checks. Column 2 of Table 1 also reports the number of education programs available for students at age 15 (OECD 2013, Figure IV.2.4; OECD 2016a, Figure II.5.8); the robustness section shows analysis based on the extent of track segregation using the number of tracks as a policy variable.

We further control for four variables to reduce the threat of grade-variant variables (that is, variables that change over time, as students move from grade four to grade eight) to our identification strategy. First, the measures of a country's public education expenditure per pupil in primary and secondary school (as percent of GDP per capita) are gathered by the Education Policy and Data Center (EPDC 2019) and the Knoema database (Knoema 2019). We combine both datasets to calculate the average change in primary education expenditure between 2007 and 2011 and the average change in secondary school spending between 2011 and 2015. The underlying assumption is that students who participated in PIRLS/TIMSS in 2011 were exposed to the education expenditures in primary school between 2007 and 2011, whereas PISA participating students were affected by secondary education expenditures between 2011 and 2015. Second, the definition of GDP growth follows the same logic. PIRLS/TIMSS 2011 students were exposed to the GDP growth between 2007 and 2011 while in primary school and PISA students were exposed to GDP growth between 2011 and 2015 while in secondary school. The variable was calculated based on GDP per capita (PPP current international dollars) supplied by the World Bank International Comparison Program database (2013).

Third. we use data on the share of female teachers in primary school in 2011 and in secondary school in 2015 provided by the UNESCO Institute for Statistics (2019).[6] Finally we control for differences in country-level gender inequality, using the

6 For Italy, the Netherlands, Norway, and Denmark, the female share in primary school in 2011 and for Bulgaria, the Czech Republic, Denmark, Spain, Finland, France, and Italy the female teacher share in secondary school in 2015 is missing. However, the time series of the female share for both primary and secondary school barely changed over time. Therefore, we impute the missing values in 2011 and 2015 for these countries by the average values of the years before and after the missing value.

Table 1: Early tracking and late tracking countries.

Country	Number of tracks	Earliest tracking age	Tracking	Sample		
				Baseline	Reading Big	Math Big
Austria	4	10	Early	√	√	√
Netherlands	7	12	Early	√	√	√
Belgium	4	12	Early	√	√	√
Czech republic	6	11	Early	√	√	√
Germany	4	10	Early	√	√	√
Italy	4	14	Early	√	√	√
Romania	2	14	Early	√	√	√
Slovak republic	5	11	Early	√	√	√
Slovenia	3	14	Early	√	√	√
Hungary	3	11	Early	√	√	√
Denmark	1	16	Late	√	√	√
Finland	1	16	Late	√	√	√
Ireland	4	15	Late	√	√	√
Portugal	3	15	Late	√	√	√
Lithuania	5	15	Late	√	√	√
Spain	1	16	Late	√	√	√
UK	1	16	Late	√	√	√
Norway	1	16	Late	√	√	√
Poland	1	16	Late	√	√	√
Sweden	1	16	Late	√	√	√
Bulgaria	3	13	Early		√	
Australia	1	16	Late		√	√
Colombia	2	15	Late		√	
France	3	15	Late		√	
Indonesia	1	15	Late		√	
Israel	2	15	Late		√	
New Zealand	1	16	Late		√	√
United States	1	16	Late		√	√
Japan	2	15	Late			√
Chile	1	16	Late			√

Source: OECD (2013 Figure IV.2.4, p. 78), OECD (2016a Figure II.5.8, p. 167) and UNESCO Institute for Statistics (2013).

Overall Global Gender Gap Index from 2011 to 2015 published by the World Economic Forum (Hausmann et al. 2011; Hausmann et al. 2015).[7]

7 The index is based on 14 indicators that measure gender inequality in economic participation and opportunity, educational attainment, health and survival, and political empowerment. Examples for such indicators are wage equality for women and men for similar work, female enrollment rates over male enrollment rates at different educational stages, and the share of women in parliament. The higher the index, the smaller a country's gender-based gaps in access to resources and opportunities.

Table 2: Summary statistics, reading.

Variable	Early tracking country		Late tracking country	
	Primary school	Secondary school	Primary school	Secondary school
Share Female	49.0 %	49.4 %	49.3 %	49.3 %
Average Score Female	498.5	505.5	511.8	528.4
Average Score Male	484.8	480.8	491.4	499.1
Share Female Teacher	90.2 %	66.8 %	81.1 %	66.1 %
Average Age	10.3	15.8	10.3	15.8
Share Books 1	11.4 %	15.1 %	8.9 %	12.6 %
Share Books 2	25.5 %	16.5 %	23.9 %	16.0 %
Share Books 3	35.9 %	30.9 %	37.2 %	31.8 %
Share Books 4	15.1 %	16.5 %	17.0 %	17.8 %
Share Books 5	12.2 %	20.9 %	13.0 %	21.9 %
Share Father Primary Education	20.3 %	12.4 %	20.7 %	18.3 %
Share Father Secondary Education	58.4 %	51.3 %	51.0 %	37.8 %
Share Father Tertiary Education	21.3 %	36.3 %	28.3 %	43.9 %
Share Mother Primary Education	19.1 %	13.3 %	15.2 %	15.1 %
Share Mother Secondary Education	58.1 %	49.6 %	47.2 %	33.9 %
Share Mother Tertiary Education	22.8 %	37.1 %	37.6 %	51.0 %
Average Minutes Reading	421.85	190.94	362.13	220.1
Female Teacher Share	90 %	67 %	81 %	66 %
GDP growth	1.1 %	1.4 %	0.9 %	2.1 %
Δ Educational expenditures	−1.0 %	−0.6 %	2.6 %	−3.1 %
Gender Index	70.6	72.8	76.8	77.8

Tables 2 and 3 show summary statistics from the datasets on reading and math, respectively. The four variables of the country-specific data is the same in both tables. The female share of teachers is higher in primary school than in secondary school, and it is higher in early tracking countries. GDP growth at the time students were in secondary school was much stronger in late tracking countries. Changes in educational spending also differed greatly across early versus late tracking countries. Late tracking countries have higher levels of gender equality. These differences in major social and economic indicators highlight the importance of including these factors in our empirical model.

Table 3: Summary statistics math.

Variable	Early tracking country		Late tracking country	
	Primary school	Secondary school	Primary school	Secondary school
Share Female	49.4 %	49.4 %	49.2 %	49.3 %
Average Score Female	495.1	494.6	499.3	504.0
Average Score Male	504.4	506.4	506.6	510.7
Average Age	10.3	15.8	10.3	15.8
Share Books 1	11.1 %	15.1 %	9.3 %	12.6 %
Share Books 2	26.4 %	16.5 %	24.8 %	16.0 %
Share Books 3	36.0 %	30.9 %	36.9 %	31.8 %
Share Books 4	15.0 %	16.5 %	16.6 %	17.8 %
Share Books 5	11.5 %	20.9 %	12.4 %	21.9 %
Share Father Primary Education	21.0 %	12.4 %	21.8 %	18.3 %
Share Father Secondary Education	58.8 %	51.3 %	52.9 %	37.8 %
Share Father Tertiary Education	20.2 %	36.3 %	25.3 %	43.9 %
Share Mother Primary Education	20.3 %	13.3 %	16.2 %	15.1 %
Share Mother Secondary Education	58.2 %	49.6 %	50.6 %	33.9 %
Share Mother Tertiary Education	21.5 %	37.1 %	33.1 %	51.0 %
Average Minutes Math	267.2	179.58	266.52	210.11
Female Teacher Share	90 %	67 %	81 %	66 %
GDP growth	1.1 %	1.4 %	0.9 %	2.1 %
Δ Educational expenditures	−1.0 %	−0.6 %	2.6 %	−3.1 %
Gender Index	70.6	72.8	76.8	77.8

The reading data in Table 2 show that parents in late tracking countries are higher educated than those in early tracking countries, and families in late tracking countries have more books in the home. Students spend more class time on reading in primary school than in secondary school; this is true in both early and late tracking countries, though the change is smaller in late tracking countries. Similar patterns can be seen in the math data in Table 3. Parental education, especially mother's education, is significantly higher in late tracking countries. Households in late tracking countries have more books. Secondary school classes in late tracking countries spend more time on math than comparable classes in early tracking countries, though as in reading, the difference between primary and secondary school is smaller in late tracking countries. These structural differences even in

individual-level background characteristics across early versus late tracking countries show the importance of comparing students with similar characteristics to get at the effects of the early tracking policy on their performance.

Without controlling for personal-and country-level background characteristics, we can see changes in the average gender-specific achievement gap across grades in early versus late tracking countries in Tables 2 and 3, as well as in Figure 1. The average gender achievement gap in Figure 1 is defined as the average test score of the lower performing sex minus the average score of the higher performing sex.[8] Therefore, a negative slope represents an increasing gender gap between primary and secondary school, whereas a positive slope indicates a decreasing gender gap. The negative slope in reading is steeper in early tracking countries than in late tracking countries. The figure gives a visualization to the descriptive statistics in Table 2. In the table we see that the gender gap in reading in late tracking countries

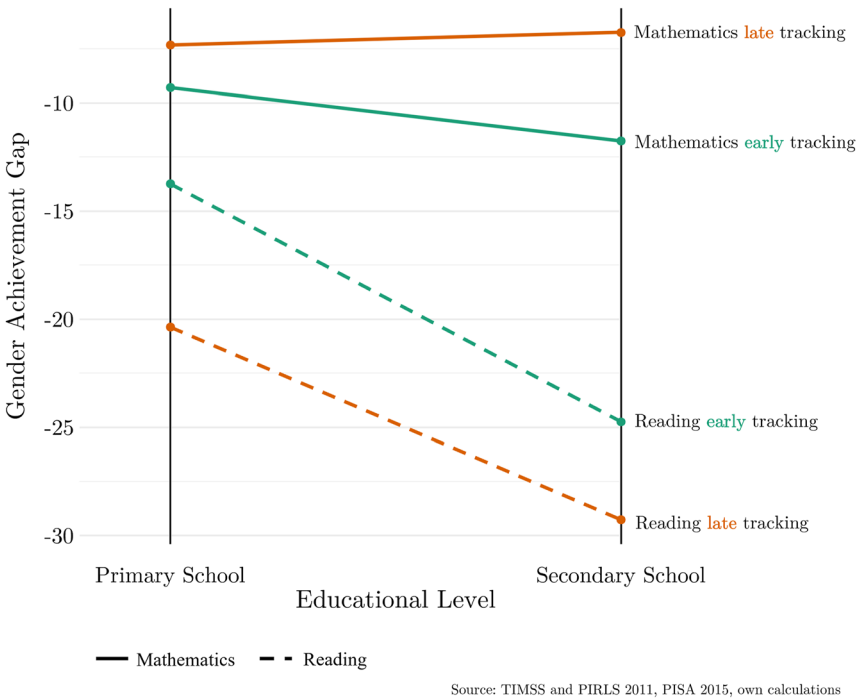

Figure 1: Average gender achievement gap in early and late tracking countries. Note: The gender gap is calculated as the difference in scores between the weaker achieving gender (girls for math, boys for reading) minus the stronger achieving gender, thus always giving a negative result.

8 Figure 1 was calculated based on the 20 countries included in the main models.

in primary school (21 points) is higher (28 points) in secondary school. The gap in early tracking countries starts at a lower point in primary school (14 points) but increases more dramatically to be much larger in secondary schools, to 26 points. For math, late tracking countries achieve a narrowing of the gender gap between primary and secondary school (the slope is positive and the gap goes from eight to seven points, as shown in Table 2), while the gender gap in early tracking countries increases slightly across school levels, from 9 to 11 points). It is interesting to note that the gender gap in reading in primary school is stronger in late tracking countries than it is in early tracking countries. We do not have an explanation for this fact but note that it highlights the importance of comparing gaps between primary school and secondary school to identify the *effect* of tracking policies, which occur after primary school; we need to account for any existing differences before the implementation of the policy.

4 Main Results

Table 4 shows the DiD results for reading and mathematics. The coefficient corresponding to the triple interaction term "Secondary × Early Tracking × Male/Female" is the main parameter of interest (β_1 in equation (1)). A negative sign implies a larger increase in the gender gap between primary and secondary school in early versus late tracking countries. The models include country fixed effects, driving the R^2 up so high.

The most important finding can be read from the top line of the table: early educational tracking increases the gender gap in reading scores, while the effect on math scores is smaller and statistically insignificant. In other words, tracking students into more specialized secondary schools increases the gender gap in reading, but does not affect the gender gap in math scores. For reading, early tracking leads to a 7.4 point higher gap in test scores, meaning that early tracking increases the gender gap in test achievement by 7.4 % of one standard deviation. This is a large impact, considering that the average gap is 25 points at the end of lower secondary school (as seen in Table 2); early tracking accounts for more than 25 % of the gap.

The DiD findings are consistent with the summary statistics given in Figure 1 and Tables 2 and 3. For reading, we can see in Table 2 that boys' scores improved from primary school to secondary school in late tracking countries (achieving 491 points in primary school and 499 in secondary school), but went down from primary school to secondary school in early tracking countries (485 to 481 points, respectively). Girls' scores increased from primary to secondary school in both early and late tracking countries, but more so in late tracking countries. Figure 1 also shows that the gender gap increased more dramatically in early tracking countries. Thus, the DiD results

Table 4: Predictions of reading and math scores, main model.

	Dependent variable:		
	Reading Score		**Mathematics Score**
Secondary × Early Tracking × Male	−7.443*** (2.488)	Secondary × Early Tracking × Female	−2.078 (3.005)
Secondary × Male	−16.166*** (2.001)	Secondary × Female	2.852 (2.871)
Secondary × Early Tracking	−0.055 (2.918)	Secondary × Early Tracking	−6.230 (4.109)
Early Tracking × Male	6.842*** (1.888)	Early Tracking × Female	−0.935 (2.268)
Secondary	−182.233*** (28.666)	Secondary	−176.928*** (28.904)
Male	91.374*** (17.241)	Female	−66.930*** (16.519)
Age	1.105 (1.240)	Age	1.432 (1.545)
Age × Secondary	14.109*** (2.090)	Age × Secondary	12.354*** (2.122)
Books 2	40.932*** (1.852)	Books 2	40.364*** (2.449)
Books 3	66.926*** (1.912)	Books 3	69.031*** (2.396)
Books 4	84.700*** (1.957)	Books 4	85.532*** (2.539)
Books 5	86.807*** (2.186)	Books 5	85.148*** (2.681)
Secondary × Books 2	−6.321*** (2.186)	Secondary × Books 2	−9.225*** (2.720)
Secondary × Books 3	−2.608 (2.346)	Secondary × Books 3	−5.047* (2.757)
Secondary × Books 4	5.818** (2.454)	Secondary × Books 4	4.602 (2.938)
Secondary × Books 5	23.756*** (2.570)	Secondary × Books 5	25.915*** (3.128)
Father Secondary Educ.	11.321*** (1.232)	Father Secondary Educ.	8.466*** (1.491)
Father Tertiary Educ.	33.775*** (1.431)	Father Tertiary Educ.	28.271*** (1.640)
Secondary × Father Sec. Educ.	−4.082** (1.710)	Secondary × Father Sec. Educ.	1.029 (1.903)
Secondary × Father Tert. Educ.	−15.321*** (1.923)	Secondary × Father Tert. Educ.	−9.197*** (2.129)
Mother Secondary Educ.	21.508*** (1.468)	Mother Secondary Educ.	20.762*** (1.789)
Mother Tertiary Educ.	44.747*** (1.562)	Mother Tertiary Educ.	45.265*** (1.926)
Secondary × Mother Sec. Educ	−8.434*** (1.813)	Secondary × Mother Sec. Educ.	−6.201*** (2.288)
Secondary × Mother Tert. Educ	−20.684*** (1.957)	Secondary × Mother Tert. Educ.	−21.315*** (2.531)
Minutes Spent Reading	−0.004 (0.006)		
Male × Minutes Reading	−0.040*** (0.007)		
		Minutes Spent in Math	0.070*** (0.016)
		Female × Minutes in Math	−0.063*** (0.017)
Gender Index	2.324***(0.461)	Gender Index	1.721*** (0.558)
Male × Gender Index	−0.770*** (0.158)	Female × Gender Index	0.480*** (0.139)
Female Teacher Share	1.394*** (0.183)	Female Teacher Share	−0.108 (0.268)
Male × Female Teacher Share	−0.615*** (0.095)	Female × Female Teacher Share	0.240** (0.114)
Δ Educ. Exp.	−1.872*** (0.226)	Δ Educ. Exp.	−0.463 (0.346)
Male × Δ Educ. Exp.	0.188 (0.158)	Female × Δ Educ. Exp.	−0.157 (0.202)
GDP growth	−4.290*** (0.525)	GDP growth	−5.308*** (0.601)
Male × GDP growth	0.930*** (0.325)	Female × GDP growth	−0.716* (0.383)

Table 4: (continued)

	Dependent variable:	
	Reading Score	**Mathematics Score**
Country fixed effects	Yes	Yes
R squared	0.97	0.97
Country Observations	20	20
Early tracking countries	10	10
Observations	232,246	226,875

Notes: Standard errors were clustered on school level; significance levels are $^{*}p<0.1$, $^{**}p<0.05$, $^{***}p<0.01$; provided senate weights were used and modified such that each country carries a weight of one; for the dependent variable "Reading Score" and "Mathematics Score" plausible values were used and standardized at mean 500 and standard deviation 100 for each survey separately.

showing that early tracking increases the gender gap in reading is not surprising. The results for math are also consistent with summary statistics in Figure 1 and Table 3, in that there were not large changes in math test scores for boys or girls between primary and secondary school in either early or late tracking countries. Thus, it is no wonder that the DiD estimate reveals no effect of early tracking on the gender gap in test scores.

Aside from the main results, three sets of other results from Table 4 are worth highlighting. First, in the school setting, a greater share of female teachers corresponds to a greater gender gap (in favor of girls) for reading and a higher share of female teachers helps to mitigate the gender gap in math scores (as seen by the positive and statistically significant coefficient on the interaction between the female teacher share and being a female student). Next, surprisingly, spending more class time on reading and math seems to be less helpful (math) or even harmful (reading) to the scores of the weaker gender in the respective subjects - though the coefficients are very small and thus perhaps not very meaningful. One explanation for the negative coefficient on the interaction term with gender in reading might be reverse causation: classes with larger gender gaps spend more time on that subject later on. For math, more class time on the subject helps boys but not girls.

Second, at home, having more books at home corresponds with higher test scores in both reading and math – an unsurprising finding. Having three or more bookcases at home (variables "Books 4" and "Books 5) is particularly helpful for secondary school students. Similarly, having more highly educated parents corresponds with higher test scores; the coefficients are larger for mother's education than father's education. Interestingly, the impact of parental education on descendant test scores is stronger in

primary school than in secondary school, as can be seen by the negative coefficients on the interaction terms between parental education and secondary school.

Third, the overall economic situation in the country is also related to test scores, and sometimes in gender-specific ways. Greater per-pupil expenditures are associated with slightly *lower* reading scores, which is surprising and concerning, though the coefficient is relatively small and again might be due to reverse causality (schools spend more in response to poor performance). Similarly, GDP growth is associated with lower test scores. One potential explanation for these findings is that the change in both variables was calculated between 2007 and 2015. During this period, the financial crisis led to major economic decline. Since the countries with the strongest educational performance in the dataset were most affected by the economic crisis,[9] a temporary negative effect of growth in GDP and educational expenditures on educational achievement might be conceivable. Finally, we see that greater gender equality in the country is associated with higher test scores in both math and reading for both boys and girls. In both subjects, it benefits the test scores of girls more than boys.

There are two potentially surprising findings in Table 4. The first is the coefficient on "male" in the model predicting reading scores: it is positive, very large, and statistically significant. This is counter-intuitive because indeed, boys have *lower* scores on reading tests, as seen in Figure 1 and Table 2. This coefficient can be explained by the interaction terms and the time-varying country-specific variables included in the model. Table A1 in the appendix shows how the coefficient on "male" changes as we add more variables to the model. In the first column with no control variables, we see that the coefficient on "male" is negative and statistically significant, as expected. As we add control variables and interaction terms between the control variables, this variable changes in sign and magnitude. In the third model in Table A1, we include the gender equality index variable and its interaction with the male dummy variable. With these inclusions, the coefficient on male alone becomes large and statistically significant. This result occurs because the value of the gender equality index is between 0 and 100, and as Tables 2 and 3 show, its average value lies between 70.6 and 77.8. Thus, the coefficient "gender equality index × male" (−0.743) times the actual value of the gender equality index (a number between 70 and 80), added to the stand-alone male coefficient gives a total male "effect" that is much smaller and much more in line with expectations. If we further consider the value of

9 For instance, Finland is by far the highest educational performance country in the analysis. Finland's average economic growth was 0.3 % between 2007 and 2015, whereas the mean growth for all other baseline countries together was 1.3 % for the same period. Other educationally high-performing countries, such as the Netherlands or Denmark, recorded below-average GDP growth in this period as well.

the coefficient on "female teacher share × male" (−0.536) times the actual value of the female teacher share (on average, a number between 67 and 90), then the overall male "effect" is indeed negative, es expected.

The second surprising result in Table 4 is the coefficient on secondary school, which is strongly negative and statistically significant in the model predicting reading scores and the model predicting math scores. Again, Tables A1 and A2 show the development of the coefficient on this variable as we include more variables in the specification. In the first column of both tables, without any controls, we see the expected positive coefficient on secondary: test scores are, on average, higher in secondary school, and this is more true for reading (Table A1) than math (Table A2) – results in line with the summary statistics in Tables 2 and 3. When we add students' background characteristics in the second column of Tables A1 and A2, the coefficient on secondary becomes strongly negative. Again, this result can be explained by the inclusion of other control variables in the model, and their interaction with secondary school status. The base level for age, for example, is zero, though all students in the data are over 10. Interacting the actual value of this variable with the coefficient on "secondary × age" (which is between 12 and 14.1 in each model in Tables A1 and A2) gives a large positive number; combining this with the stand-alone coefficient on "secondary" gets us closer to the expected (positive) result. The same thinking is true for all other variables interacted with secondary school status; the number of books at home is particularly powerful. Combining the coefficients on the variables interacted with secondary school with the actual value of the variables and adding these together shows a positive relationship between secondary school and test scores.

In sum, early tracking is found to increase gender gaps in reading (7.3 % of one standard deviation). In math, there is a statistically insignificant relationship between early tracking and the gender gap in test scores. These results are in contrast to some studies in the literature; given the differences in the empirical strategies employed, the divergence in findings is not particularly surprising. The findings in van Hek et al. (2019) suggest that early tracking narrows the gender gap in reading; our results are more in line with Hermann and Kopasz (2019), who find early tracking to increase the reading gap. For math test scores, our results are in line with Bodovski et al. (2020), who find no statistically significant relationship between early tracking and the gender gap.

5 Robustness Checks

Most of our robustness checks are similar to those performed by Hanushek and Wößmann (2006), Ruhose and Guido (2016), and Hermann and Kopasz (2019). The

tables in this section only report the parameter of interest (β_1), but the estimates are based on the complete model used in the main analysis.

5.1 Definition of Early Tracking

Our first set of robustness analyses investigates whether the main results are driven by the definition of early tracking (thus far, the cut-off was tracking before the age of 15). In the first column of Tables 5 and 6, this definition is replaced by the number of education programs available for students at the age of 15. The number of available tracks varies between one educational program for very late tracking countries and up to seven educational programs for early tracking countries (see Table 1). The measure of the number of tracks available is an indication of the differentiation of a country's school system. A negative coefficient on the variable "number of tracks" indicates that greater differentiation before age 15 leads to higher gender differences in test achievement. Table 5 reveals that this is exactly the case for reading. In mathematics, the coefficient in Table 6 is close to zero; more differentiation does not affect the gender gap in math scores.

Next, instead of strictly separating early and late tracking countries with a particular cut-off age as in the main analysis, it is also possible to use the actual age of first tracking, measured by a continuous variable. As shown in Table 1, the tracking age varies between 10 in Austria and Germany and 16 in Northern European and

Table 5: Different definitions of early tracking in reading.

	Dependent variable: Reading Score (main model DiD estimate: −7.443***)			
	Early tracking = no. of tracks [1–7]	**Actual age of first tracking**	**Without very early tracking countries**	**Early tracking = tracking before age 13**
Secondary × Early Tracking × Male	−1.872*** (0.601)	1.112** (0.514)	−5.830* (2.427)	−6.530*** (2.427)
Country fixed effects	Yes	Yes	Yes	Yes
Country observations	20	20	15	20
Early tracking countries	10	10	5	7
Observations	232,246	232,246	175,427	232,246

Notes: Standard errors were clustered on school level; significance levels are $^{*}p<0.1$, $^{**}p<0.05$, $^{***}p<0.01$; provided senate weights were used and modified such that each country carries a weight of one; for the dependent variable "Reading Score" plausible values were used and standardized at mean 500 and standard deviation 100 for each survey separately. Very early is defined as tracking before the age of 12.

Table 6: Different definitions for early tracking in mathematics.

	Dependent variable: Mathematics score (full model DiD estimate: 2.078)			
	Early tracking = no. of tracks [1–7]	**Actual age of first tracking**	**Without very early tracking countries**	**Early tracking = tracking before age 13**
Secondary × Early Tracking × Female	0.383 (0.871)	0.733 (0.617)	−1.034 (4.424)	−0.353 (2.969)
Country fixed effects	Yes	Yes	Yes	Yes
Country observations	20	20	15	20
Early tracking countries	10	10	5	7
Observations	226,875	226,875	163,879	226,875

Notes: Standard errors were clustered on school level; significance levels are $^{*}p<0.1$, $^{**}p<0.05$, $^{***}p<0.01$; provided senate weights were used and modified such that each country carries a weight of one; for the dependent variable "Mathematics Score" plausible values were used and standardized at mean 500 and standard deviation 100 for each survey separately. Very early is defined as tracking before the age of 12.

some Southern European countries. This test examines whether small changes (of one year) in the tracking age change the effect of early tracking on the gender gap in test scores. Results of this exercise are shown in the second column of Tables 5 and 6. A positive coefficient indicates that earlier tracking increases the gender gap in test scores. This is the case for reading scores: for both subjects, every one year increase in the age of first tracking leads to a smaller negative effect of early tracking on the gender gap in test achievement. For math test scores, the effect of tracking at an earlier age defined in this way remains statistically insignificant.

In the third robustness check, we examine whether the main results are driven by countries with very early tracking. To do so, we remove the five countries whose first tracking occurs at age 10 or 11 from the sample. The results are given in the third column of Tables 5 and 6. The results are consistent with the main findings. Even without very early tracking countries in the sample, early tracking increases the gender gap in reading, and it has no statistically significant effect on the gender gap in math scores. Finally, we check the robustness of the exact cutoff age used in the main analysis. Here we change the cut-off of early versus late tracking from 15 to 13. Using this definition, there is still a strong effect of early tracking on the gender gap in reading. The coefficient for math remains statistically insignificant.

Taken together, the results of these analyses show that the main findings are not driven by the definition of "early" tracking or the exact age at which the tracking occurs.

5.2 Additional Countries

The main analysis uses data from 20 European countries. Using data from countries in other regions gives us the opportunity to check the main model's sensitivity to a broader range of institutional contexts. Similar results for this extended model would be an indication that the results for the 20 European countries in the main results have international external validity. The additional countries and their tracking ages can be seen in the lower portion of Table 1. Countries added for the reading and math models are shown in the column "Reading Big" and "Math Big," respectively. The choice of countries was based solely on the availability of variables used in the main model. There are eight additional countries for the big sample for reading (seven of which are late tracking) and five additional countries for the big sample for math (all of which are late tracking). The results are presented in Table 7.[10] Adding the new countries results in only minor changes in the parameter of interest for the reading results. However, for math, the coefficient becomes large and statistically significant. We thus have evidence that the effect of early tracking on the gender gap in math test scores depends on the countries included in the sample.

A second test addresses the role of any one country on determining the results. Minor differences for the effect of early tracking depending on the sub-sample of countries used can be expected. It is, however, important to ensure that the main

Table 7: Results using the big sample in reading and mathematics.

	Big Sample (main model DiD estimate: −7.443 and −2.078)***	
	Reading	**Mathematics**
Secondary × Early Tracking × Gender	−6.172*** (2.028)	−4.216** (2.020)
Country fixed effects	Yes	Yes
R^2	0.974	0.976
Country observations	28	25
Early tracking countries	11	10
Observations	322,275	321,450

Notes: Standard errors were clustered on school level; significance levels are *p<0.1, **p<0.05, ***p<0.01; senate weights were used such that each country carries a weight of one; for the dependent variable plausible values were used and standardized at mean 500 and standard deviation 100 for each survey separately. Due to missing data in the big mathematics model, parents' education was dropped.

10 Due to missing data in the big sample math model, mother's and father's education was dropped.

model result is not solely driven by certain unobservable characteristics of one particular country. Following Ruhose and Guido (2016) to rule out a single country's relevance for the main results, we perform a piece-wise deletion of one country at a time and a re-estimation of the main model. Table 8 gives these results. The first column shows which country was excluded from the analysis. The table is read as follows: If Austria is excluded from the sample, i.e. the main model is estimated for nine early and 10 late tracking countries, the parameter of interest is −7.132 in reading and −1.125 in mathematics. If only the Czech Republic is excluded, the values change to −6.347 in reading and to −3.316 in mathematics.

Table 8: Impact of dropping single countries in reading and mathematics.

Dependent variable: Reading Score (main model DiD estimate: −7.443***)			
Excluded country	**Secondary × Early Tracking × Male**	**Excluded country**	**Secondary × Early Tracking × Male**
Austria	−7.132*** (2.527)	Belgium	−8.085*** (2.572)
Czech republic	−6.347** (2.567)	Germany	−8.146*** (2.604)
Denmark	−6.958** (2.571)	Spain	−7.996*** (2.565)
Finland	−8.204*** (2.565)	Great Britain	−5.902** (2.618)
Hungary	−7.475*** (2.540)	Ireland	−9.429*** (2.553)
Italy	−7.937*** (2.527)	Lithuania	−7.682*** (2.527)
Netherlands	−6.967*** (2.579)	Norway	−9.159*** (2.535)
Poland	−7.018*** (2.597)	Portugal	−5.458** (2.547)
Romania	−6.245** (2.461)	Slovakia	−7.540*** (2.649)
Slovenia	−6.352** (2.604)	Sweden	−8.784*** (2.497)
Dependent variable: Mathematics Score (full model DiD estimate: −2.078)			
Excluded country	**Secondary × Early Tracking × Female**	**Excluded country**	**Secondary × Early Tracking × Female**
Austria	−1.125 (3.133)	Belgium	−2.180 (3.207)
Czech republic	−3.316 (2.950)	Germany	−1.266 (3.215)
Denmark	−2.112 (3.184)	Spain	−3.141 (3.022)
Finland	−0.459 (3.058)	Great Britain	−3.414 (3.224)
Hungary	−2.536 (3.133)	Ireland	−0.640 (2.983)
Italy	−1.178 (3.163)	Lithuania	−3.008 (3.1015)
Netherlands	−2.264 (3.007)	Norway	−0.257 (3.094)
Poland	−2.535 (3.057)	Portugal	−2.920 (3.117)
Romania	−2.911 (3.024)	Slovakia	−1.829 (3.004)
Slovenia	−2.291 (2.987)	Sweden	−1.581 (3.105)

Notes: Standard errors were clustered on school level; significant levels are $^{*}p<0.1$, $^{**}p<0.05$, $^{***}p<0.01$; provided senate weights were used and modified such that each country carries a weight of one; for the dependent variables plausible values were used and standardized at mean 500 and standard deviation 100 for each survey separately; coefficients are based on the full models.

In reading, results are statistically significant no matter which country is excluded. It is interesting to note that when Northern European countries are excluded, the effect of early tracking on gender differences in reading increases (that is, the coefficient raises above the baseline level of 7.4 points). These results emerge because all Northern European countries are late tracking countries with relatively large gender differences in reading scores. In mathematics, regardless of which single European country is left out of the analysis, the coefficient remains statistically insignificant.

5.3 Dataset Timing

In this section, we check the robustness of our results using earlier waves of the data and using TIMSS data from both grade four and grade eight. These exercises allow us to check if the effect of early tracking is stable over time, and if our models produce results similar to other studies using the same data. To use older data, we must drop four countries in the reading analysis and up to seven countries in the math analysis due to missing data. When we do the analysis comparing TIMSS data from grade four and grade eight, we only have data on two early tracking countries.

The first analysis in this section changes the combination of dataset years used. The main analysis matches PIRLS/TIMSS 2011 with PISA 2015 data, which allowed us to observe students at different schooling levels who were born in the same year, thus eliminating potential birth cohort effects. We deal with any potential biasing time effects by interacting the change in educational expenditures, GDP growth, the share of female teachers, and the gender index with the gender of the pupil. Matching two datasets of the same year eliminates the possibility of any time effects, but the existence of biasing birth cohort effects may exist. Matching PIRLS/TIMSS 2011 data with PISA 2012 data removes most time effects but introduces potential cohort effects. For this exercise, we do not include the country controls used in the main analysis (that is, changes in GDP growth, educational expenditures, share of female teachers, and gender inequality index over time) because those over-time changes are irrelevant when we compare datasets from the same year or from just one year apart.

The results of matching PIRLS/TIMSS 2011 with PISA 2012 data can be found in the first column of Tables 9 and 10. They show that the negative effect of early tracking on gender differences in reading achievement are slightly smaller but similar to the main results. The results for math using these two datasets are very similar to the results in the main model: a negative but small and statistically insignificant effect of early tracking.

Table 9: Different data matching in reading.

	*Dependent variable: Reading Score (main model DiD estimate: −7.443*** *)*		
Data	**PIRLS 11**	**PIRLS 06**	**PIRLS 06**
Match	**PISA 12**	**PISA 12**	**PISA 12**
Secondary × Early Tracking × Male	−5.154** (2.178)	−12.362*** (2.594)	−11.312*** (2.677)
Country fixed effects	Yes	Yes	Yes
R^2	0.969	0.969	0.970
Country observations	20	16	16
Early tracking countries	10	9	9
Observations	261,242	213,019	205,614

Notes: Standard errors were clustered on school level; significance levels are *p<0.1, **p<0.05, ***p<0.01; senate weights were used such that each country carries a weight of one; for the dependent variable "Reading Score" plausible values were used and standardized at mean 500 and standard deviation 100 for each survey separately. The second PIRLS 06 and PISA 12 match additionally accounts for parents' country of birth.

Table 10: Different data matching in mathematics.

	Dependent variable: Mathematics Score (main model DiD estimate: −2.078)			
Data	**TIMSS 11**	**TIMSS 07**	**TIMSS 07**	**TIMSS grade 4**
Match	**PISA 12**	**PISA 12**	**PISA 12**	**TIMSS grade 8**
Secondary × Early Tracking × Female	−1.098 (3.202)	−1.025 (3.212)	−2.359 (2.577)	−2.359 (2.577)
Country fixed effects	Yes	Yes	Yes	Yes
R^2	0.97	0.97	0.97	0.97
Country observations	20	13	13	10
Early tracking countries	10	5	5	2
Observations	255,871	151,912	146,695	107,598

Notes: Standard errors were clustered on school level; significance levels are *p<0.1, **p<0.05, ***p<0.01; senate weights were used such that each country carries a weight of one; for the dependent variable "Mathematics Score" plausible values were used and standardized at mean 500 and standard deviation 100 for each survey separately. The second TIMSS 07 and PISA 12 match additionally accounts for parents' country of birth.

Our second exercise looks at how the effect of early tracking may have changed over time. Theoretically, the effect of early tracking on gender differences should not change significantly over small periods of time. To test this, we calculate the DiD estimate by pooling primary school data from 2006 PIRLS (for reading) and 2007 TIMSS (for math) with secondary school data from PISA 2012; results are given in the second column in Tables 9 and 10. In these earlier data, we do not have information on four countries that were in the main analysis (Czech Republic, Finland, Ireland,

and Portugal). The coefficient on the effect of early tracking in reading gender gaps was over 12 points in the earlier data. When pairing the earlier data for the math analysis, we must drop seven countries (Romania, Belgium, Finland, Spain, Ireland, Portugal, and Poland) from the analysis due to data constraints. The result is parallel to the main findings: a small, negative, and statistically insignificant effect of early tracking.

Third, the impossibility of controlling for migrant status in the main models might have led to biased results. Unlike the 2011 datasets, TIMSS 2007 and PIRLS 2006 measure parents' country of birth.[11] In the third column of Tables 9 and 10, this variable was added to the model in column two. When comparing columns two and three, any biased results in relation to the student's migrant status can be ruled out; the migration variable does not change the estimate in any meaningful way.

Finally, we consider that TIMSS data are not only available at grade four but also measures students' test scores four years later, at grade eight. Matching TIMSS grade four with TIMSS grade eight data, instead of only using PISA data at the secondary level, has two advantages. First, since PISA and TIMSS don't measure exactly the same, we can test whether results are sensitive to the particular secondary school data used. The second advantage is that PISA and TIMSS grade eight surveys are carried out for sightly different student ages. In the PISA study, students are usually slightly older. We did not use TIMSS grade four and grade eight data as our main model because TIMSS has data on math but not reading. In addition, for our set of control variables, only a limited number of countries are available (only two early tracking countries and eight late tracking countries).[12] Results using data from TIMSS 2011 grade four and TIMSS 2015 grade eight are presented in column four of Table 10. The coefficient of interest is slightly larger than the coefficient in the main model, but it too is statistically insignificant.

In sum, the robustness checks show that the effect of early tracking on gender differences in reading is consistent across different years and countries included in the analysis. We conclude that the main results for the effect of early tracking on the gender gap in reading scores are robust. In mathematics, the main results are statistically insignificant and remain so throughout almost all robustness checks. The one exception is when we include five non-European late-tracking countries; their inclusion leads to a statistically significant coefficient measuring the effect of early tracking. Otherwise, there is overwhelming evidence that for the 20 European countries in the sample, early tracking does not lead to an increase in the gender gap in math test scores.

---

11 Our migration background variable in these specifications is equal to one if at least one parent was born outside the test country.

12 All countries in the main model were used for estimation, but many countries are missing data on key control variables.

6 Conclusions

In this paper, we investigated the impact of early educational tracking on the widening gender gap in educational performance between primary and secondary school. We estimated a difference-in-differences model using cross-country variation in the age of first educational tracking as a policy variable. We used large-scale international test studies at the primary and secondary school level for reading and math. Applying these data allowed us to compare boys' and girls' test scores between early and late tracking countries in secondary school, conditional on gender differences that already existed in primary school. By applying a cross-country DiD model taking into account a wide range of control variables absent in previous literature and all characteristics of a complex sample design, this study could produce causal estimates of the effect of early tracking on the change in the gender gap in test scores between primary and secondary school.

The results for reading indicate that early tracking increases gender differences between primary and secondary school. A wide range of robustness checks confirm the results. Our empirical results on the gender gap in reading scores are in line with the the first stream of the maturity hypothesis, which predicted stronger gender gaps in reading for countries with early tracking instead of late tracking. The reading results are also consistent with our gender roles hypothesis, which predicted that early tracking widens the gender gap both in reading and mathematics. For math scores, almost across the board, there is no evidence that early tracking affects the gender gap in test scores. The sole exception is when we include five non-European late-tracking countries in the sample. Overall, the results do not give evidence to conclude that early tracking impacts the gender gap in math scores.

The findings for reading are in line with the results in Hermann and Kopasz (2019), while van Hek et al. (2019) found the opposite. The latter used non-causal methods and is thus not comparable. As reproduced in this study, the results for the effect of early tracking on the gender gap in mathematics are sensitive to several details of the analysis. Using less recent data, Bedard and Cho (2010) concluded that early tracking related to worse performance for girls in mathematics. Hermann and Kopasz (2019), however, have shown early tracking to be positively related to girls' performance in mathematics. Bodovski et al. (2020), like us, found a statistically insignificant relationship between early tracking and gender gaps in mathematics.

The key conclusion drawn from this paper is that early tracking is likely to contribute to a widening gender gap between primary and secondary school in reading. The reading results give clear implications: early tracking increases gender differences in academic achievement.

Appendix

Table A1: Prediction of reading score, main results.

	Dependent variable: Reading Score			
	(1)	**(2)**	**(3)**	**(4)**
Secondary × Early Tracking × Male	−2.215 (2.594)	−5.993** (2.562)	−6.752*** (2.541)	−7.443*** (2.488)
Secondary × Male	−9.103*** (1.662)	−12.388*** (1.792)	−14.861*** (1.860)	−16.166*** (2.001)
Secondary × Early Tracking	−9.745*** (2.484)	−0.752 (2.758)	−1.207 (2.772)	−0.055 (2.918)
Early Tracking × Male	6.851*** (1.709)	8.533*** (1.762)	5.892*** (1.918)	6.842*** (1.888)
Secondary	16.978*** (1.568)	−156.307*** (29.138)	−164.289*** (28.846)	−182.233*** (28.665)
Male	−20.398*** (1.208)	9.821 (6.559)	84.073*** (16.647)	91.374*** (17.241)
Age		3.223*** (1.231)	2.736** (1.224)	1.105 (1.240)
Secondary × Age		12.229*** (2.131)	12.958*** (2.106)	14.109*** (2.090)
Books 2		40.912*** (1.840)	40.907*** (1.841)	40.932*** (1.852)
Books 3		67.102*** (1.902)	66.998*** (1.903)	66.926*** (1.912)
Books 4		85.083*** (1.945)	84.876*** (1.949)	84.700*** (1.957)
Books 5		87.442*** (2.180)	87.175*** (2.184)	86.807*** (2.186)
Secondary × Books 2		−6.135*** (2.180)	−6.227*** (2.179)	−6.321*** (2.186)
Secondary × Books 3		−2.830 (2.339)	−2.725 (2.338)	−2.608 (2.346)
Secondary × Books 4		5.085** (2.442)	5.375** (2.445)	5.818** (2.454)
Secondary × Books 5		22.406*** (2.570)	22.786*** (2.576)	23.756*** (2.571)
Father Secondary Educ.		10.231*** (1.240)	10.264*** (1.240)	11.321*** (1.232)
Father Tertiary Educ.		33.129*** (1.438)	33.090*** (1.438)	33.775*** (1.431)
Secondary × Father Secondary Educ.		−3.038* (1.722)	−3.068* (1.723)	−4.082** (1.710)
Secondary × Father Tertiary Educ.		−14.702*** (1.934)	−14.612*** (1.934)	−15.321*** (1.923)
Mother Secondary Educ.		20.980*** (1.469)	21.071*** (1.470)	21.508*** (1.468)
Mother Tertiary Educ.		45.124*** (1.571)	45.292*** (1.572)	44.747*** (1.562)
Secondary × Mother Secondary Educ.		−8.105*** (1.815)	−8.241*** (1.815)	−8.434*** (1.813)
Secondary × Mother Tertiary Educ.		−21.246*** (1.971)	−21.644*** (1.974)	−20.684*** (1.957)
Minutes Reading		−0.002 (0.006)	−0.004 (0.006)	−0.004 (0.006)
Male × Min. reading		−0.045*** (0.007)	−0.043*** (0.007)	−0.040*** (0.007)
Female Teacher Share		1.311*** (0.183)	1.518*** (0.182)	1.394*** (0.183)
Male × Female Teacher Share		−0.325*** (0.080)	−0.536*** (0.090)	−0.615*** (0.095)

Table A1: (continued)

	Dependent variable: Reading Score			
	(1)	**(2)**	**(3)**	**(4)**
Gender Index			1.765*** (0.454)	2.324*** (0.461)
Male × Gender Index			−0.743*** (0.154)	−0.770*** (0.158)
Δ Educ. Exp.				−1.872*** (0.226)
Male × Δ Educ. Exp.				0.188 (0.158)
GDP growth				−4.290*** (0.525)
Male × GDP growth				0.930*** (0.325)
Country fixed effects	Yes	Yes	Yes	Yes
R squared	0.964	0.97	0.97	0.97
Country observations	20	20	20	20
Early tracking countries	10	10	10	10
Observations	237,169	232,246	232,246	232,246

Notes: Standard errors clustered on school level; significance levels are *p<0.1, **p<0.05, ***p<0.01; provided senate weights were used and modified such that each country carries a weight of one; for the dependent variable "Reading Score" plausible values were used and standardized at mean 500 and standard deviation 100 for each survey separately.

Table A2: Prediction of math score, main results.

	Dependent variable: Mathematics Score			
	(1)	**(2)**	**(3)**	**(4)**
Secondary × Early Tracking × Female	−2.735 (2.660)	−3.165 (2.955)	−3.077 (2.954)	−2.078 (3.005)
Secondary × Female	0.904 (1.864)	0.273 (2.723)	1.969 (2.742)	2.852 (2.871)
Secondar × Early Tracking	−2.279 (2.765)	2.874 (3.658)	1.489 (3.739)	−6.230 (4.109)
Early Tracking × Female	−2.318 (1.793)	−1.802 (2.214)	0.235 (2.253)	−0.935 (2.268)
Secondary	3.984**	−155.360***	−161.697***	−176.928***
	(1.847)	(28.879)	(28.766)	(28.906)
Female	−7.476***	−15.023* (8.162)	−61.748***	−66.930***
	(1.312)		(15.885)	(16.518)
Age		3.669** (1.546)	3.392** (1.532)	1.432 (1.545)
Secondary × Age		10.368*** (2.120)	10.794*** (2.107)	12.354*** (2.122)
Books 2		40.058*** (2.438)	40.073*** (2.442)	40.364*** (2.449)
Books 3		68.788*** (2.379)	68.721*** (2.385)	69.031*** (2.396)
Books 4		85.470*** (2.521)	85.329*** (2.532)	85.532*** (2.539)
Books 5		85.403*** (2.672)	85.209*** (2.682)	85.148*** (2.681)
Secondary × Books 2		−8.619*** (2.711)	−8.675*** (2.712)	−9.225*** (2.720)
Secondary × Books 3		−4.514* (2.739)	−4.409 (2.749)	−5.047* (2.757)
Secondary × Books 4		4.772 (2.911)	5.025* (2.934)	4.602 (2.938)
Secondary × Books 5		25.524*** (3.097)	25.864*** (3.127)	25.915*** (3.128)
Father Secondary Educ.		7.623*** (1.479)	7.678*** (1.482)	8.466*** (1.491)

Table A2: (continued)

	Dependent variable: Mathematics Score			
	(1)	**(2)**	**(3)**	**(4)**
Father Tertiary Educ.		27.877*** (1.642)	27.901*** (1.644)	28.271*** (1.640)
Secondary × Father Secondary Educ.		1.804 (1.911)	1.738 (1.912)	1.029 (1.903)
Secondary × Father Tertiary Educ.		−9.069*** (2.145)	−9.104*** (2.144)	−9.197*** (2.128)
Mother Secondary Educ.		20.513*** (1.783)	20.607*** (1.786)	20.762*** (1.789)
Mother Tertiary Educ.		45.406*** (1.916)	45.703*** (1.927)	45.265*** (1.926)
Secondary × Mother Secondary Educ.		−6.438*** (2.277)	−6.619*** (2.284)	−6.201*** (2.288)
Secondary × Mother Tertiary Educ.		−22.025*** (2.501)	−22.612*** (2.533)	−21.315*** (2.531)
Minutes Math		0.059*** (0.016)	0.061*** (0.016)	0.070*** (0.016)
Female × Min. Math		−0.047*** (0.016)	−0.050*** (0.016)	−0.063*** (0.017)
Female Teacher Share		0.394 (0.266)	0.334 (0.266)	−0.108 (0.268)
Male × Female teacher share		0.045 (0.097)	0.178* (0.105)	0.240** (0.113)
Gender Index			0.832 (0.550)	1.721*** (0.558)
Female × Gender Index			0.465*** (0.136)	0.480*** (0.139)
Δ Educ. Exp.				−0.463 (0.346)
Female × Δ Educ. Exp.				−0.157 (0.202)
GDP growth				−5.308*** (0.601)
Female × GDP growth				−0.716* (0.384)
Country fixed effects	Yes	Yes	Yes	Yes
R squared	0.965	0.97	0.97	0.97
Country observations	20	20	20	20
Early tracking countries	10	10	10	10
Observations	246,710	226,875	226,875	226,875

Notes: Standard errors were clustered on school level; significant levels are $^{*}p<0.1$, $^{**}p<0.05$, $^{***}p<0.01$; provided senate weights were used and modified such that each country carries a weight of one; for the dependent variable "Mathematics Score" plausible values were used and standardized at mean 500 and standard deviation 100 for each survey separately.

References

Altermatt, E.R. and Pomerantz, E.M. (2005). The implications of having high-achieving versus low-achieving friends: a longitudinal analysis. Soc. Dev. 14: 61–81.

Ammermüller, A. (2005). *Educational opportunities and the role of institutions*. Tech. rep. 05–44, Mannheim, Germany, ZEW Discussion Papers.

Anderson, V.A., Anderson P., Northam E., Jacobs R., and Catroppa C. (2001). Development of executive functions through late childhood and adolescence in an Australian sample. Dev. Neuropsychol. 20: 385–406.

Baye, A. and Monseur, C. (2016). Gender differences in variability and extreme scores in an international context. Large-Scale Assessments. Educ. 4: 1–16.

Bedard, K. and Cho, I. (2010). Early gender test score gaps across OECD countries. Econ. Educ. Rev. 29: 348–363.

Bodovski, K., Munoz, I., Byun, S.Y., and Chykina, V. (2020). Do education system characteristics moderate the socioeconomic, gender and immigrant gaps in math and science achievement? Int. J. Sociol. Educ. 9: 122–154.

Connor, C.M., Morrison F.J., Fishman B., Crowe E.C., Al Otaiba S., Schatschneider C. (2013). A longitudinal cluster-randomized controlled study on the accumulating effects of individualized literacy instruction on students' reading from first through third grade. Psychol. Sci. 24: 1408–1419.

Daniel, W.A., Jr., Duke, P.M., Carlsmith, J.M., Jennings, D., Martin, J.A., Dornbusch, S.M., Gross, R.T., and Siegel-Gorelick, B. (1982). Educational correlates of early and late sexual maturation in adolescence. J. Pediatr. 100: 633–637.

Delaruelle, K., van de Werfhorst, H., and Bracke, P. (2019). Do comprehensive school reforms impact the health of early school leavers? Results of a comparative difference-in-difference design. Soc. Sci. Med. 239: 1–10.

Diefenbach, H. and Klein, M. (2002). Bringing boys back in. Soziale Ungleichheit zwischen den Geschlechtern im Bildungssystem zuungunsten von Jungen am Beispiel der Sekundarschulabschlüsse. Z. für Padagogik 48: 938–958.

DiPrete, T.A. and Jennings, J.L. (2012). Social and behavioral skills and the gender gap in early educational achievement. Soc. Sci. Res. 41: 1–15.

Dubas, J.S., Graber, J.A., and Petersen, A.C. (1991). The effects of pubertal development on achievement during adolescence. Am. J. Educ. 99: 444–460.

Eckert, P. (1989). *Jocks and burnouts: social categories and identity in the high school*. Teachers College Press, New York.

Else-Quest, N.M., Hyde, J.S., and Linn, M.C. (2010). Cross-national patterns of gender differences in mathematics: a meta-analysis. Psychol. Bull. 136: 103–127.

EPDC (2019). *Education policy and data center. Education expenditure. Public education expenditure per pupil (% of GDP per capita)*, Available at: https://www.epdc.org/topic/education-expenditure (Accessed 20 August 2019).

Fryer, Jr., Roland, G., and Levitt, S.D. (2010). An empirical analysis of the gender gap in mathematics. Am. Econ. J. Appl. Econ. 2: 210–240.

Gallagher, A.M. and Kaufman, J.C. (2004). *Gender differences in mathematics: an integrative psychological approach*. Cambridge University Press, New York.

Guiso, L., Monte F., Sapienza P., and Zingales L. (2008). Culture, gender, and math. Science 320: 1164–1165.

Hadjar, A., Krolak-Schwerdt, S., Priem, K., and Glock, S. (2014). Gender and educational achievement. Educ. Res. 56: 117–125.

Hanushek, E.A., Kain, J.F., Markman, J.M., and Rivkin, S.G. (2003). Does peer ability affect student achievement? J. Appl. Econom. 18: 527–544.

Hanushek, E.A. and Wößmann, L. (2006). Does educational tracking affect performance and inequality? Differences-in-differences evidence across countries. Econ. J. 116: C63–C76.

Hausmann, R., Tyson, L.D., and Zahidi, S. (2011). The global gender gap report 2011. World Economic Forum, Geneva, Switzerland.

Hausmann, R., Tyson, L.D., and Zahidi, S. (2015). The global gender gap report 2015, 10th Anniversary ed World Economic Forum, Geneva, Switzerland.

Hermann, Z. and Kopasz, M. (2019). Educational policies and the gender gap in test scores: a cross-country analysis. Res. Pap. Educ. 36: 1–22.

Hörstermann, T., Krolak-Schwerdt, S., and Fischbach, A. (2010). Die kognitive Repräsentation von Schülertypen bei angehenden Lehrkräften-Eine typologische Analyse. Schweiz. Z. Bildungswissenschaften 32: 143–158.

Huang, M.H. (2009). Classroom homogeneity and the distribution of student math performance: a country-level fixed-effects analysis. Soc. Sci. Res. 38: 781–791.

Hyde, J.S. and Mertz, J.E. (2009). Gender, culture, and mathematics performance. Proc. Natl. Acad. Sci. USA 106: 8801–8807.

IEA (2019a). *PIRLS 2011 international database*, Available at: https://timssandpirls.bc.edu/pirls2011/international-database.html (Accessed 21 May 2019).

IEA (2019b). *TIMSS 2011 international database*, Available at: https://timssandpirls.bc.edu/timss2011/international-database.html (Accessed 21 May 2019).

Jürges, H. and Schneider, K. (2011). Why young boys stumble: early tracking, age and gender bias in the German school system. Ger. Econ. Rev. 12: 371–394.

Keulers, E.H., Evers E.A.T., Stiers P., and Jolles J. (2010). Age, sex, and pubertal phase influence mentalizing about emotions and actions in adolescents. Dev. Neuropsychol. 35: 555–569.

Knoema (2019). *World data atlas. Country profiles. Education. Expenditure on education*, Available at: https://knoema.de/ (Accessed 22 May 2019).

Langen, A.v, Rekers-Mombarg, L., and Dekkers, H. (2008). Mathematics and science choice following introduction of compulsory study profiles into Dutch secondary education. Br. Educ. Res. J. 34: 733–745.

Lee, V.E. and Bryk, A.S. (1986). Effects of single-sex secondary schools on student achievement and attitudes. J. Educ. Psychol. 78: 1–381.

Legewie, J. and DiPrete, T.A. (2012). School context and the gender gap in educational achievement. Am. Socio. Rev. 77: 463–485.

Li, Q. (1999). Teachers' beliefs and gender differences in mathematics: a review. Educ. Res. 41: 63–76.

Ma, X. (2001). Participation in advanced mathematics: do expectation and influence of students, peers, teachers, and parents matter? Contemp. Educ. Psychol. 26: 132–146.

Maurin, E. and McNally, S. (2007). Educational effects of widening access to the academic track: a natural experiment. Tech. rep. 2596. Bonn, Germany, IZA discussion paper.

Montt, G. (2011). Cross-national differences in educational achievement inequality. Sociol. Educ. 84: 49–68.

Mullis, I.V.S., Martin, M.O., and Loveless, T. (2015). *20 years of TIMSS: international trends in mathematics and science achievement, curriculum, and instruction*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College, Chestnut Hill, MA.

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

OECD (2009b). *PISA data analysis manual: SPSS*, 2nd ed. Paris, France: OECD Publishing, pp. 1–475, Available at: https://www.oecd-ilibrary.org/content/publication/9789264056275-en.

OECD (2013). *PISA 2012 results: what makes schools successful? resources, policies and practices (volume IV)*.

OECD (2016a). *PISA 2015 results (Volume II)*, pp. 1–468, Available at: https://www.oecd-ilibrary.org/content/publication/9789264267510-en.

OECD (2016b). *Results: excellence and equity in education*, Vol. I. Paris, France.

OECD (2019). *About - PISA*, Available at: http://www.oecd.org/pisa/aboutpisa/ (Accessed 21 May 2019).

Pekkarinen, T. (2008). Gender differences in educational attainment: evidence on the role of tracking from a Finnish quasi-experiment. Scand. J. Econ. 110: 807–825.

Pinxten, M., De Fraine B., Van Den Noortgate W., Van Damme J., Anumendem D. (2012). Educational choice in secondary school in Flanders: the relative impact of occupational interests on option choice. Educ. Res. Eval. 18: 541–569.

Robertson, D. and James S. (2003). Do peer groups matter? Peer group versus schooling effects on academic attainment. Economica 70: 31–53.

Ruhose, J. and Guido, S. (2016). Does early educational tracking increase migrant-native achievement gaps? Differences-in-differences evidence across countries. Econ. Educ. Rev. 52: 134–154.

Schindler, S. and Bittmann, F. (2021). Diversion or inclusion? Alternative routes to higher education eligibility and inequality in educational attainment in Germany. Eur. Socio Rev. 37: 972–986.

Statistik Austria (2017). *Bildung in Zahlen 2015/16 - Schlüsselindikatoren und Analysen*. Statistik Austria, Vienna, Austria.

Stoet, G. and Geary, D.C. (2015). Sex differences in academic achievement are not related to political, economic, or social equality. Intelligence 48: 137–151.

Tanner, J.M. (1990). *Foetus into man: physical growth from conception to maturity*. Harvard University Press, Cambridge, MA, pp. 1–288.

Traini, C., Kleinert, C, and Bittmann, F. (2021). How does exposure to a different school track influence learning progress? Explaining scissor effects by track in Germany. Res. Soc. Stratif. Mobil. 76: 100625.

Turner, S.E. and Bowen, W.G. (1999). Choice of major: the changing (unchanging) gender gap. ILR Review 52: 289–313.

UNESCO Institute for Statistics (2013). *Unesco's world data on wducation. Profiles of national educational systems*, Available at: https://www.ecml.at/News/TabId/643/ArtMID/2666/ArticleID/25/Unesco's-World-Data-on-Education.aspx (Accessed 21 May 2019).

UNESCO Institute for Statistics (2019). *Unesco's world data on education*, Available at: https://data.worldbank.org/indicator/se.prm.tchr.fe.zs (Accessed 22 May 2019).

Van de Werfhorst, H.G. (2017). Vocational and academic education and political engagement: the importance of the educational institutional structure. Comp. Educ. Rev. 61: 111–140.

van Hek, M., Buchmann, C., and Kraaykamp, G. (2019). Educational systems and gender differences in reading: a comparative multilevel analysis. Eur. Socio Rev. 35: 169–186.

World Bank International Comparison Program Database (2013). *Worldbank international comparison program database*, Available at: https://data.worldbank.org/indicator/NY.GDP.PCAP.PP.CD?end=2015&start=2011&year_high_desc=true (Accessed 22 May 2019).

Ziegler, A., Kuhn, C., and Heller, K.A. (1998). Implizite Theorien von gymnasialen Mathematik-und Physiklehrkräften zu geschlechtsspezifischer Begabung und Motivation. Psychol. Beiträge 40: 271–287.