Routledge
Taylor & Francis Group

Research Papers in Education

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/rred20

Educational policies and the gender gap in test scores: a cross-country analysis

Zoltán Hermann & Marianna Kopasz

To cite this article: Zoltán Hermann & Marianna Kopasz (2021) Educational policies and the gender gap in test scores: a cross-country analysis, Research Papers in Education, 36:4, 461-482, DOI: 10.1080/02671522.2019.1678065

To link to this article: https://doi.org/10.1080/02671522.2019.1678065

© 2019 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

Published online: 16 Oct 2019.

View supplementary material

Submit your article to this journal

Article views: 6029

View related articles

View Crossmark data

Citing articles: 3 View citing articles

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=rred20

RESEARCH PAPERS IN EDUCATION
2021, VOL. 36, NO. 4, 461–482
https://doi.org/10.1080/02671522.2019.1678065

OPEN ACCESS

Educational policies and the gender gap in test scores: a cross-country analysis

Zoltán Hermann[a,b] and Marianna Kopasz[c]

[a]Centre for Economic and Regional Studies of the Hungarian Academy of Sciences, Institute of Economics, Budapest, Hungary; [b]Centre for Labour Economics, Corvinus University of Budapest, Budapest, Hungary; [c]Centre for Social Sciences of the Hungarian Academy of Sciences, Institute for Political Science, Budapest, Hungary

ABSTRACT

Girls tend to outperform boys in reading tests, while they usually lag behind boys in mathematics. However, the size of the gender gap varies to a great extent between countries. While the existing figliterature explains these differences as being mainly due to cultural factors, this paper explores whether this cross-country variation is related to educational policies like tracking, grade retention, and individualised teaching practices. The gender test score gap is analysed in mathematics, reading and science using the PISA 2012 dataset. Multilevel models are used in the estimation. The results suggest that the extent of the gender gap is indeed associated with certain characteristics of the various education systems. First, applying a difference-in-differences estimation method, it was found that early tracking has a direct effect on the gender gap in test scores, in favour of girls. Second, suggestive evidence shows that more student-oriented teaching practices also benefit girls relative to boys, both between and within countries, and within schools. Finally, grade retention is correlated with the gender gap, though there is further evidence suggesting that this correlation is very unlikely to represent a causal effect.

ARTICLE HISTORY
Received 19 February 2019
Accepted 5 October 2019

KEYWORDS
Gender gap; tracking; grade retention; teaching practice; PISA; multilevel model

1. Introduction

Gender differences in educational achievement are a well-known phenomenon. In most countries, boys score higher in mathematics tests, while girls tend to do better in reading. This is a concern for policy-makers since, in spite of the increase in higher education enrolment, girls are still heavily underrepresented in STEM (Science, Technology, Engineering and Mathematics) fields (see, e.g. OECD 2015). At the same time, in recent decades, girls have been closing the test score gap in subjects traditionally favouring boys, like mathematics and science, while extending their advantage in reading literacy (Marks 2008; Baye and Monseur 2016). In recent decades female educational attainment has also risen rapidly, and in most of the developed countries more women than men obtain a higher education degree. These trends are often

CONTACT Zoltán Hermann hermann.zoltan@krtk.mta.hu Institute of Economics, Centre for Economic and Regional Studies, Tóth Kálmán u. 4., Budapest 1097, Hungary

Supplemental data for this article can be accessed here.

© 2019 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

regarded as a symptom of the so-called 'boy crisis'. The widening gender gap in educational attainment in the era of labour market polarisation can lead to increasing inequalities and decreasing labour market participation among men (Pekkarinen 2012). Nevertheless, research on the test score gap still focuses mainly on girls' disadvantage in mathematics (Stoet and Geary 2015).

It is also well documented that the gender gap in test scores varies remarkably between countries (Schnepf 2004; Marks 2008; Fryer and Levitt 2010). In some countries, girls lag behind boys in mathematics by a considerable margin, while in others they are on average at par. Similarly, in some cases, girls outdistance boys in reading to a considerable degree, while at the other extreme, only a narrow gap can be observed. The comparative gender gap research seeking explanations for these differences almost exclusively focuses on the role of cultural factors, social norms and female participation in the labour market and politics (see, e.g. Penner 2008; Guiso et al. 2008; Else-Quest, Hyde, and Linn 2010). Cross-country differences in the gender gap are for the most part linked to gender inequalities in society and, to a lesser extent, to gender role attitudes and beliefs.

However, the existing evidence is mixed. Some studies conclude that there is a link between gender inequality measures (e.g. women's participation in politics, the labour market, etc., or composite indicators of gender inequality) and the gender test score gap (e.g. Riegle-Crumb 2005; Else-Quest, Hyde, and Linn 2010; Guiso et al. 2008; González de San Román and De la Rica Goiricelaya 2012). Others challenge this conclusion (Fryer and Levitt 2010; Stoet and Geary 2015).

At the same time, it seems obvious that schools play a decisive role in mediating the effects of societal and cultural factors. Indirect evidence also suggests that schools indeed affect the gender gap in achievement. The gender gap in mathematics abilities opens up after children enter school (Fryer and Levitt 2010), and in school this gap increases with age (i.e. from primary school to secondary school) both in reading, mathematics and science (Baye and Monseur 2016). Moreover, schooling seems to have heterogeneous effects across gender. Boys appear to benefit more from higher school quality (Autor et al. 2016) and perceived teacher quality (Hochweber and Vieluf 2018) than girls.

As the gender gap seems to be formed in schools, it is natural to assume that the specific characteristics of various education systems affect the cross-country differences in the gender gap (Ayalon and Livneh 2013). Prior research shows that the gender gaps in reading and mathematics are highly correlated at the country level (Van Langen, Bosker, and Dekkers 2006; Guiso et al. 2008; Marks 2008; González de San Román and De la Rica Goiricelaya 2012). In other words, where girls have a larger advantage in reading over boys, they also tend to have a smaller disadvantage in mathematics. This suggests that cross-country differences in the gender gap are not determined by educational policies specific to a given subject (i.e. the curriculum in mathematics) (Marks 2008), but rather, that broader educational institutions and policies are at work.

However, the effect of national educational policies has hardly been addressed in the literature, and the few existing studies focus on the homogeneity of school systems. Higher degrees of standardisation and integration in the education system were found to be associated with a higher relative performance by girls (Ayalon and Livneh 2013; Van Langen, Bosker, and Dekkers 2006). At the same time, early tracking seems to benefit boys (Van Hek, Buchmann, and Kraaykamp 2019; Bedard and Cho 2010). Altogether, prior evidence is scarce, and confined to a limited set of educational institutions.

This study seeks to contribute to filling this void by analysing the relationship between educational policies (or characteristics of the education systems) and gender differences in educational performance from a cross-country perspective. More specifically, the focus is on three policies: early tracking,[1] the extensive use of grade retention, and the incidence of individualised teaching practices. These are very different features, but all of these are among the key educational policies that education systems use to manage the heterogeneity of the student population (Mons 2004, 2007). In this respect these policies have similar functions. In another framework, proposed by OECD (2013b), these educational policies are conceived as forms of vertical and horizontal stratification.

This study aims to explore whether and to what extent these educational policies (early tracking, grade retention, and the use of individualised teaching) contribute to the cross-country variation in the gender gap in test scores. These question are addressed using a two-stage empirical strategy. First, the association between the gender gap and the educational policies is explored. In this stage, multilevel regression models are employed including all three educational policy variables at the same time. Second further evidence on the effect of each policy variable separately is provided. The effect of grade retention is tested indirectly, by comparing its effects in different parts of the performance distribution. For early tracking, its direct effect is identified using a difference-in-differences strategy. Finally, suggestive evidence is offered concerning the effects of individualised teaching relying on within-country and within-school variation in teaching practices.

In this paper, data from the 2012 wave of OECD’s PISA programme is used. For the analysis of early tracking, these are supplemented by the IEA’s PIRLS and TIMSS data from 2006 and 2007. PISA provides data on students’ test scores in mathematics, reading, and science for more than sixty countries, including all OECD member countries. Here, data for a single cross-section are used, as the cross-country patterns of the gender gap hardly change over time.

The contribution made here to the literature is threefold. First, the evidence on the relationship between educational policies and gender differences in student achievement is scarce. This study explores the effects of three educational policies. Two of these, grade retention and teaching practice, have not been analysed in this context before. Second, most of the evidence on cross-country differences in the gender gap is descriptive, confined to correlations. In this paper, the causal effect of early tracking is identified, and suggestive evidence provided in the case of the other two policy variables. Third, despite the fact that the disadvantage of boys in reading is a growing concern, the vast majority of the gender gap literature focuses on mathematics only. In contrast, this paper covers three fields of competence. Analysing mathematics, science and reading within one study, it is possible to shed light on which policies favour boys or girls, and in which field of competence.

The remainder of the paper is structured as follows. In the next section, the context of the research is outlined and a review provided of previous research investigating the effects of educational policies on gender differences. Section 3 describes the data and the estimation methods used. In Section 4 the results are presented. Finally, Section 5 draws conclusions.

2. Theoretical background and related literature

Prior evidence suggests that the cross-country differences of the gender test score gap are related to the characteristics of educational systems. However, this literature looked at only a few educational institutions. Van Langen, Bosker, and Dekkers (2006) examine the degree of integration for the educational systems (measured by such factors as grade and track differentiation, number of tracks, socioeconomic, gender and immigrant segregation, and quality differences) in relation to the gender gap. Integrated educational systems are found to be more favourable to the achievement of girls (in mathematics, science and reading) than differentiated ones.

Ayalon and Livneh (2013) address the gender effects of educational standardisation. They report a significant level of association between the degree of standardisation and gender difference in mathematics test scores. According to this, a higher degree of standardisation (i.e. the use of national examinations and the higher uniformity in time devoted to various mathematics topics) is linked to a reduced gender gap in mathematics test scores. An apparently different conclusion is reached for reading performance by Van Hek, Buchmann, and Kraaykamp (2019). They demonstrate that a higher degree of standardisation is associated with a larger gender gap (favouring girls) in reading achievement. However, boys outperform girls in mathematics, while lagging behind in reading. Hence, both studies suggest that standardisation provides relative benefits for girls.

Van Hek, Buchmann, and Kraaykamp (2019) also explored the relationship between the gender achievement gap and early tracking in the cross-country context. She found that the gender gap in reading scores is smaller in countries that track students at an early age, i.e. early tracking provides relative benefits to boys.

In a similar study Bedard and Cho (2010) analysed correlations between educational institutions and the gender gaps in mathematics and science across developed countries. Their results show that tracking countries tend to have larger gender gaps at 8th grade level. Further, and more interestingly, the same is true in grade 4, long before formal tracking occurs in most countries.

These studies are motivated by the research into the effect of educational institutions on inequality of opportunity. This thread in the literature asks how educational institutions shape the effect of family background on students' educational achievement. In international comparison, the key question is why some countries are more successful in offsetting socio-economic inequalities and ensuring greater equality of opportunity in schools than others. Early tracking and standardisation are key themes in this literature (Van de Werfhorst 2015; Hanushek and Woessmann 2010). It is natural to ask whether the mechanisms behind educational policies affecting socioeconomic inequalities in achievement also affect gender inequality.

This paper seeks to contribute to the literature on the cross-country differences in the gender test score gap by exploring the effect of other educational policies. The analysis is informed by the theoretical framework elaborated by Mons (2004, 2007, see also Dupriez et al. 2008); she starts from the observation that in response to students' diverse abilities, school systems use different policies to manage heterogeneity in the student population. She identifies four key educational policies developed to deal with student heterogeneity: tracking, ability grouping, grade retention, and individualised teaching practices. Tracking and ability grouping allow for student sorting based on ability and motivation, resulting in more homogeneous classes. In theory, this leaves room for adjusting the level and content of education to fit students' needs better. Grade retention decreases heterogeneity within classes by holding back students who cannot meet minimum achievement standards. Finally, the use of individualised teaching practices implies allocating additional teacher time and attention to help low-achievers. Mons argues that though the countries studied rely on a mix of these measures, one of them tended to become predominant in most cases. Based on the particular policy mix implemented in a given country, Mons (2004, 2007) and Janmaat and Mons (2011) distinguish between four models of heterogeneity management. Selective school systems use early tracking, while comprehensive school systems rely on either ability grouping within and across schools or frequent grade retention or individualised teaching practices to deal with student heterogeneity.

Though these policies have similar functions, the empirical evidence suggests that they are not equally effective. There are large differences in equality of opportunity across countries (Schütz, Ursprung, and Woessmann 2008) and also across the types identified by Mons (Dupriez, Dumay, and Vause 2008; Castejón and Zancajo 2015).

Seeking to explain cross-country differences in the gender gap the effects of these policies on the gender gap are explored. Only three of the four policies are investigated, ignoring ability grouping, as no reliable and comparable measure is available on that at the country level. Prior research on the relationships between educational policies and the gender gap is only available with regard to early tracking.

The empirical evidence largely confirms that early tracking strengthens the influence of parental background on students' educational achievement, as tracking has a detrimental effect on low-achievers (Hanushek and Woessmann 2006; Schütz, Ursprung, and Woessmann 2008; Bol and van de Werfhorst 2013; Lavrijsen and Nicaise 2015). At the same time, in tracking regimes boys tend to be more often placed in lower tracks than girls (Van Hek, Buchmann, and Kraaykamp 2019). This gender inequality in track placement may hinder the educational performance of boys through different mechanisms (Van Hek, Buchmann, and Kraaykamp 2019). One such mechanism is that classroom homogeneity may hamper achievement in lower tracks (Huang 2009; Van Hek, Buchmann, and Kraaykamp 2019). Another mechanism that may hinder boys' performance is related to the prevailing school norms that differ between school tracks. Due to their lower track placement, boys are less likely to be exposed to norms supportive of academic performance (Scheeren, Van de Werfhorst, and Bol 2018; Van Hek, Buchmann, and Kraaykamp 2019).

Some direct evidence on the effect of early tracking on the gender gap in educational attainment is also available. Pekkarinen (2008), analysing the effect of a Finnish comprehensive school reform of the 1970s, reports that the shift from a selective school system to a comprehensive one had a positive effect on girls' probability of choosing the academic track later, whereas this effect was slightly negative for boys. These findings suggest that postponing tracking favours girls. Scheeren, Van de Werfhorst, and Bol (2018) test a similar hypothesis by using microdata from the European Social Survey and longitudinal data on tracking age reforms for 21 European countries from 1929 until 2000. The main finding of their study is that later tracking improves girls' completed years of education relative to boys'.

In the cases of the use of grade retention and individualised teaching practices, the empirical evidence is mostly limited to studies analysing the effects of these on individual educational attainment (i.e. individual level analyses). However, prior research lends support to the assumption that these policies may matter, as it suggests that boys and girls are affected differently by these policies.

With regard to teaching practices, the empirical literature most often contrasts two types: lecture-style teaching and teaching based on problem-solving. The former is often associated with more traditional, didactic teacher-centred teaching styles, while the letter is associated with more modern, interactive, student-oriented teaching styles (Schwerdt and Wuppermann 2011). The latter can be conceived as similar to individualised teaching in the terminology of Mons (2007). Analysing Spanish data, Hidalgo-Cabrillana and Lopez-Mayan (2015) conclude that modern teaching practices are associated with better student performance, especially in reading, while traditional practices if anything, are disadvantageous. These effects differ according to gender: girls gain from modern practices and lose from traditional ones, while boys do not benefit from any particular teaching style. Korbel and Paulus (2017) investigated the effect of teaching practices on non-cognitive skills using Czech data and found that the effects are different by gender.

The effect of grade retention on student achievement at the individual level is often hotly debated. The empirical evidence is mixed and fairly controversial. The general conclusion is that grade retention has either no effect or has a negative impact on student performance (Jimerson, Anderson, and Whipple 2002; Jimerson et al. 2006; Martin 2009; Manacorda 2012). At the student level, grade retention effect is most often found not to differ by gender (Martin 2009; Ikeda and García 2014). In contrast, Morrison and No (2007) report a more detrimental impact on boys. At the same time, it is often observed that boys stand a higher risk of repeating a grade than girls (Jimerson et al. 2006; Martin 2009). This implies that if repeating a grade has a direct effect on student achievement, boys are affected to a greater extent by grade retention policies overall, due to their higher exposure to this policy.

In this paper we examine the question of whether and how these educational policies affect the gender gap in achievement. Based on a review of the literature, it appears that the relationship between the gender test score gap and use of grade retention/individualised teaching practices are under-researched areas. As for these educational policies, no well-grounded hypotheses can be offered concerning their effects on the gender gap in achievement, and thus our study is of exploratory nature.

Previous research, however, enables us to formulate a hypothesis on the effect of early tracking on the gender test score gap. Based on the theoretical considerations outlined above (see Van Hek, Buchmann, and Kraaykamp 2019), we would expect that early tracking is more beneficial to girls. At the same time, the empirical research evidence so far does not support this claim (see e.g. Pekkarinen 2008; Scheeren, Van de Werfhorst, and Bol 2018; Van Hek, Buchmann, and Kraaykamp 2019). In this paper we retest this hypothesis using a different methodology (i.e. we analyse the direct effect of early tracking employing a difference-in-differences strategy).

3. Data and methods

3.1. Data

The primary dataset used in this paper is the 2012 wave of the OECD Programme for International Student Assessment (PISA). PISA is a survey of 15-year-old students in which skills in different domains are assessed: mathematics, literacy and science, with the major focus on mathematical literacy in 2012. PISA 2012 was implemented in 65 countries, including all 34 OECD member countries. We use a single wave of PISA, as the cross-country patterns of the gender gap are fairly stable over time. Moreover, measures of teaching practices in mathematics are available only for 2012.

Our final sample contains 472,074 students from 62 countries. Cyprus is not included in the available data set, while Lichtenstein is excluded due to the small number of observations. Furthermore, Taiwan is also excluded as the Global Gender Gap Index is not available.

Besides the PISA data, student achievement data from the Progress in International Reading Literacy Study (PIRLS) and the Trends in International Mathematics and Science Study (TIMSS) datasets are used in Section 5.2. PIRLS and TIMSS are standardised student achievement testing programmes similar to PISA carried out by the International Association for the Evaluation of Educational Achievement (IEA). PIRLS tests students at grade four in reading, and TIMSS tests students in mathematics and science.

3.2. Variables

The dependent variables in the analysis are mathematics, reading and science test scores, standardised within each country, so that the mean is 0, and standard deviation is 1. In this way any differences in the overall level of performance between countries are removed from the data and the gender differences are directly comparable across countries.[2]

The key variables describe educational policies at the country level. Tracking is measured as the age at which students are first tracked into different school types. Data on the age of first selection are gathered from the OECD (2013b). The tracking variable is truncated at the age of 15, since it is assumed that achievement measured at this age is not affected by tracking that occurs later.

Data on grade retention were gathered from the PISA student questionnaire. The country-level variable is measured as the share of students who have repeated a grade at least once at either the primary or secondary level. In the regressions, the natural logarithm of this variable is used, as it fits the data better.

To measure individualised teaching the index student-orientated teaching practices developed by the OECD (2013a) is used. This index was constructed using students' reports on the frequency with which, in mathematics lessons, the teacher gives different work to classmates who have difficulties learning and/or to those who can advance faster; the teacher assigns projects that require at least one week to complete; the teacher has students work in small groups to come up with a joint solution to a problem or task; and the teacher asks students to help plan classroom activities or topics (OECD 2013a). Higher values of the index indicate the more intensive use of student orientated practices. It should be noted that, though measuring teacher classroom behaviour based on students' responses may contain considerable measurement error, student-reported measures are more closely related to student achievement than those reported by teachers (Hidalgo-Cabrillana and Lopez-Mayan 2015). The questions on teaching practices in PISA 2012 refer to mathematics lessons only. This measure is used as a proxy for teaching practices in general at the country level, assuming a strong correlation between subjects. In other words, it is assumed that teaching practices reflect a general pedagogical approach and teaching culture rather than subject-specific methodological differences at the country level. At the same time, when comparing schools within countries this correlation cannot be presumed; hence, the effect of teaching practices across and within schools is analysed only in the case of mathematics.

In the analysis gender stratification is controlled for by the use of one composite index of gender inequality: the Global Gender Gap Index (GGI), prepared by the World Economic Forum (2009 data). This index is widely used in the literature (Guiso et al. 2008; Fryer and Levitt 2010) and available for most of the PISA countries. The GGI is comprised of four sub-indices which measure economic participation and opportunity, educational attainment, political empowerment, health and survival. Larger GGI values indicate a better position of women in society. To control for economic development GDP per capita (logarithmic form) is used. GDP data for 2011 are derived from World Bank data base.

In the estimated models two student characteristics are controlled for: immigrant background and parental education.

Summary statistics for the key country-level variables are displayed in Table 1. It should be noted that all the educational policy variables display a significant degree of variation from country to country.

Besides the variables used in the analysis, Table 1 also presents country-level measures of the gender gap in the three subjects. The gender gap is calculated as the weighted mean score of males minus the weighted mean score of females.

In reading, boys lag behind girls in each country by a considerable margin; the gap typically falls between standard deviations (SD) of −0.1 and −0.9. At the same time, in most countries, on average, boys outperform girls in mathematics. However, the difference tends to be smaller; while in a few countries girls are on a par with or better than boys. In the case of science the picture is mixed, with boys in some countries performing better, while in others, girls excel. The gap typically varies between −0.5 and 0.2 SD. It is important to note that the gender gaps in the three domains correlate strongly at the country level. In countries where girls have a large advantage in reading, they also tend to close the gap in mathematics and perform better than boys in science (Marks 2008).

Table 1. Descriptive statistics of country-level variables.

	N	minimum	maximum	mean	standard deviation
Gender gap (male-female)					
Maths gender gap (PISA)	62	−0.282	0.365	0.092	0.123
Reading gender gap (PISA)	62	−0.862	−0.136	−0.444	0.142
Science gender gap (PISA)	62	−0.548	0.245	−0.029	0.134
Maths gender gap (TIMSS)	27	−0.270	0.213	0.037	0.123
Reading gender gap (PIRLS)	30	−0.420	−0.054	−0.209	0.093
Science gender gap (TIMSS)	27	−0.252	0.196	0.016	0.122
Educational policy variables					
Tracking age	62	10	15	14.178	1.499
Grade retention (log)	62	−0.693	3.717	2.046	1.184
Individualised teaching	62	−0.579	1.081	0.205	0.397
Country-level controls					
Gender Gap Index	62	0.400	0.828	0.696	0.069
GDP per capita (log)	62	8.459	11.794	10.192	0.654

Variables derived from PIRLS and TIMSS are calculated for PISA countries only.

3.3. Estimation methods

We explore the effects of educational policies on the gender gap using multilevel regression models. Our baseline model similar to those used in the existing literature to estimate the effect of standardisation and early tracking (Ayalon and Livneh 2013; Van Hek, Buchmann, and Kraaykamp 2019). Test score variables are explained with individual characteristics including gender, country level variables including indicators for educational policies and interaction terms of gender and educational policies (see Appendix 1 for the detailed model specification).

In the second stage of the analysis, further evidence is provided on the effects of the three education policy variables separately, extending the model in different directions. Unfortunately, the available data does not often allow for proper identification of causal effects in a country-level analysis. Hence various empirical strategies were employed. First, an indirect implication of grade retention effects was tested by re-estimating the baseline model and comparing the results for subsamples of students over different parts of the performance distribution. Grade retention might be expected to have a direct effect only on low achievers, implying different correlations across the distribution. The direct effect of early tracking is then analysed employing a difference-in-differences strategy, augmenting the dataset with 4th graders and adding further interaction terms to the baseline model. Here the question of whether the gender achievement gap develops differently between grade 4 and age 15 in tracking and non-tracking countries is tested. Finally, suggestive evidence is provided concerning the effect of individualised teaching exploiting the variation between and within schools. Here a three-level model is employed, extending the baseline model with a school level. The exact model specifications are described at the beginning of each section.

It should be noted that only the difference-in-differences analysis of the tracking effect can be considered as a causal identification strategy per se. In the other cases, suggestive evidence is provided that is non-causal, but none the less helps to assess the effects of educational policies on the gender gap.

4. Results

4.1. Baseline model

Table 2 presents the results of the base multilevel model. As between-country gender inequality is represented by the variation of the gender slope, the key parameters of interest are the interaction terms of gender and the education policy variables. These coefficients indicate whether the presence of an educational policy on average goes together with an additional (dis)advantage to girls relative to boys, compared to countries where this policy is used to a lesser extent.

The results suggest that grade retention has the most consistent correlation with the gender gap. The interaction between the share of students who have repeated a grade and gender is statistically significant for all three subjects. The negative coefficients indicate that a higher rate of grade retention tends to be favourable to boys. In mathematics, for example, a unit increase in log grade retention (e.g. an increase from 10 to 26 percentage points in the share of grade repeaters) results in a decrease of 0.0454 SD in the female-male test score gap. In other words, on average, girls perform better relative to boys in

Table 2. Education policies and the gender test score gap: baseline model.

	maths	reading	science
	(1)	(2)	(3)
Student variables			
female	−0.306	0.623**	0.0718
	(0.214)	(0.305)	(0.248)
parental education (ref. cat.: upper secondary)			
lower secondary or below	−0.327***	−0.318***	−0.324***
	(0.0237)	(0.0249)	(0.0246)
tertiary	0.383***	0.363***	0.377***
	(0.0171)	(0.0158)	(0.0168)
immigrant background	−0.0509	−0.0587	−0.0900
	(0.0576)	(0.0538)	(0.0599)
Country variables			
log grade retention	0.0662***	0.0590***	0.0595***
	(0.0114)	(0.0128)	(0.0123)
tracking age	0.00399	0.00502	0.00215
	(0.00754)	(0.00739)	(0.00781)
student-oriented teaching	−0.0784**	−0.0540	−0.0807*
	(0.0395)	(0.0467)	(0.0442)
Gender Gap Index	−0.356*	−0.345*	−0.217
	(0.182)	(0.190)	(0.192)
log GDP per capita	−0.0737**	−0.0444	−0.0531
	(0.0351)	(0.0368)	(0.0388)
Cross-level interactions			
female X log grade retention	−0.0454***	−0.0376**	−0.0311**
	(0.0121)	(0.0149)	(0.0132)
female X tracking age	−0.00503	−0.00694	0.00103
	(0.00692)	(0.00710)	(0.00683)
female X student-oriented teaching	0.152***	0.0961	0.162***
	(0.0401)	(0.0589)	(0.0483)
female X Gender Gap Index	−0.0508	−0.0693	−0.353*
	(0.165)	(0.242)	(0.210)
female X log GDP per capita	0.0389**	0.00382	0.0229
	(0.0196)	(0.0268)	(0.0224)
constant	0.739**	0.169	0.420
	(0.377)	(0.388)	(0.413)
Random-effects parameters			
variance of country level random intercept	0.0069	0.0095	0.0089
	(0.0017)	(0.0021)	(0.0022)
variance of country level random slope (female)	0.0049	0.0132	0.0083
	(0.0019)	(0.0037)	(0.0030)
individual level residual variance	0.9348	0.8913	0.9363
	(0.0049)	(0.0057)	(0.0049)
Observations	472,074	472,074	472,074
Number of countries	62	62	62

Robust standard errors clustered at the country-level in parentheses.
***p < 0.01, **p < 0.05, *p < 0.1

countries where grade retention is less prevalent. This implies that a strict grade retention policy goes together with a larger gender gap in mathematics, as on average boys outperform girls in mathematics in most countries, and with a smaller gap in reading.

Individualised teaching also seems to matter. It is significantly associated with the gender gap in mathematics and science, and it is marginally insignificant for reading (p = 0.103). In science, for example, a 1SD increase in the index of student-oriented teaching implies an increase by 0.162SD in girls' test scores relative to boys. These results suggest that the widespread use of student-oriented teaching practices conveys more benefits to girls, especially in mathematics and science.

As opposed to grade retention and individualised teaching, tracking age appears to have no effect on the gender slope in the baseline model presented here. The coefficients are highly nonsignificant for each subject.[3]

How large is the estimated effect of individualised teaching and grade retention? In order to assess effect size, it is important to note that the standard deviation of log grade retention at the country level is about three times that of the student-oriented teaching indicator (see Table 1). Taking this into account, the two-to-five times larger coefficients of student-oriented teaching indicate an effect of similar magnitude. In other words, a one standard deviation change of log grade retention and the index of student-oriented teaching implies a similar change in the gender gap.

Overall this first set of results suggests that two of the three educational policies are associated with the gender gap at the country level. A higher frequency of grade retention tends to favour boys, while more individualised teaching practices appear to benefit girls relative to boys, especially in the case of mathematics and science. At the same time, early tracking is not associated with the gender gap.

However, it is important to emphasise that these coefficients represent country-level correlations. This is prima facie evidence, which does not necessarily represent causal effects, and thus requires further verification. In the following sections, further evidence is sought for the effects of the three policy variables, using various empirical strategies.

4.2. Grade retention

The results of our baseline model suggest that grade retention is closely related to the gender gap. The higher the share of grade repeaters in a country, the better boys perform relative to girls on average. However, it should be pointed out that interpreting this association thus, in causal terms may well be mistaken. To provide further evidence an indirect implication of grade retention effects is tested.

In most cases, students repeating a grade fail to reach a minimum standard. Cross-country differences in grade retention occur as these standards may differ between countries or because performing below standard does not necessarily incur repeating a grade in some countries. In either case, if grade retention has a direct effect on the gender gap (e.g. repeating a grade affects student performance differently by gender, or the threat of it motivates boys more than girls), its effect should be stronger on low-achievers. As high-achievers rarely repeat a grade, they are directly not affected by the retention rate.

This implication was tested by comparing grade retention effects on the gender gap measured in different parts of the test score distribution. The sample within each country was split into three groups with respect to the test score and the baseline model for the low, middle, and high-achiever groups was estimated separately. The set of independent variables in the model remain unchanged.

Table 3 gives the estimated coefficients for the education policy – female student interaction terms. The results show no marked differences across the test score distribution in the association between grade retention and the gender gap. A higher retention rate goes together with the better performance of boys relative to girls both among low- and high-achievers. For mathematics, the estimated coefficients are almost identical in the three groups. For reading and science, the coefficients slightly decrease moving

Table 3. Education policies and the gender test score gap: low-, middle- and high-achievers.

	Maths	Reading	Science
	(1)	(2)	(3)
Low-achievers			
female X log grade retention	−0.0469***	−0.0457***	−0.0370**
	(0.0122)	(0.0164)	(0.0145)
Middle-achievers			
female X log grade retention	−0.0487***	−0.0387**	−0.0358**
	(0.0136)	(0.0162)	(0.0141)
High-achievers			
female X log grade retention	−0.0479***	−0.0344**	−0.0280**
	(0.0136)	(0.0138)	(0.0135)

Each panel represents the cross-level interactions from a separate regression estimate. Model specification is identical to that in Table 2. Robust standard errors clustered at the country-level are given in parentheses.
***p < 0.01, **p < 0.05, *p < 0.1

upwards on the achievement scale, but pairwise tests of the equality of the coefficients across the groups reveal no statistically significant differences.

Comparing students with different socio-economic backgrounds instead of different parts of the test score distribution provides similar results (results available upon request).

Consequently, our indirect test does not support the existence of a direct effect; retention policies per se hardly affect the gender gap. It is more likely that retention policy is correlated with other characteristics of the education systems that influence the gender test score gap, and represents the effects of these unobserved factors in country-level regressions.

4.3. Early tracking

In this section, we investigate further the correlation between the gender gap and early tracking. The baseline model reveals no significant association here. However, if some unobserved characteristics of the education system are correlated with both early tracking and gender test score gap, the estimated effect of early tracking is biased and the true effect of early selection may be concealed.

To test the direct effects of early tracking a difference-in-differences approach was employed (see Ammermüller 2005; Waldinger 2007; Lavrijsen and Nicaise 2015). Combining PISA data with PIRLS or TIMSS datasets measuring achievement in the fourth grade provides an ideal setting, as PISA measures students after tracking has taken place in early tracking countries, while in late tracking countries there is no tracking at the age of 15.

Using the difference-in-differences method we control for the effects of grade-invariant unobserved confounding factors, and the model identifies the causal effect of early tracking on inequalities. This approach builds on the observation that early tracking should not affect student achievement in primary education, which is untracked in every country. We calculate the increase in the gender gap from grade 4 to age 15 in each country. Then we compare the average increase in the group of early- tracking countries and in the group of non-tracking countries. We assume that in the absence of early tracking the increase in the first group would be the same as it is in the non-tracking group. In other words, this assumption says that the effects of unobserved country characteristics, including educational institutions are fully captured in the gender gap in grade 4. Under this assumption, the difference between the two groups can be attributed to early tracking.

Figure 1 demonstrates this idea in the case of reading. The figure depicts the gender gap in reading test scores in primary education, measured in PIRLS 2006 for fourth graders and in secondary education, measured in PISA 2012 for the 15-year-olds. As may immediately be seen, the gender gap widens in every country, except Great Britain.

However, if the change in the gender gap from primary to secondary education in the two country groups is compared, the patterns show an interesting difference. The dashed lines in the figure represent the values for the gender gap that might be expected at the secondary level, given the value of the gender gap at the primary level. The short and long dashed lines correspond to early and late or non-tracking countries respectively. At a given level of the gender gap in primary education, girls' advantage tends to increase more in early tracking countries.

To test the direct effect of early selection formally, the PISA dataset was augmented with the PIRLS and TIMSS samples of 4th graders, and the baseline model was extended by interaction terms between gender, early tracking and an indicator variable denoting PISA students (see Appendix 2 for the detailed model specification).

Table 4 gives the estimates for the education policy–female student interaction effects. In columns 1, 3 and 5 tracking is measured with the age of selection under age 15, as before. In the other columns, a dummy variable specification is employed, as is frequently the case in the tracking literature. Non-tracking denotes countries that use a comprehensive school system or track students later than the age of 14.[4] The number of countries is about half of the full PISA sample, as here only those countries participating both in the PISA and the PIRLS or TIMSS programme at fourth grade level are included.

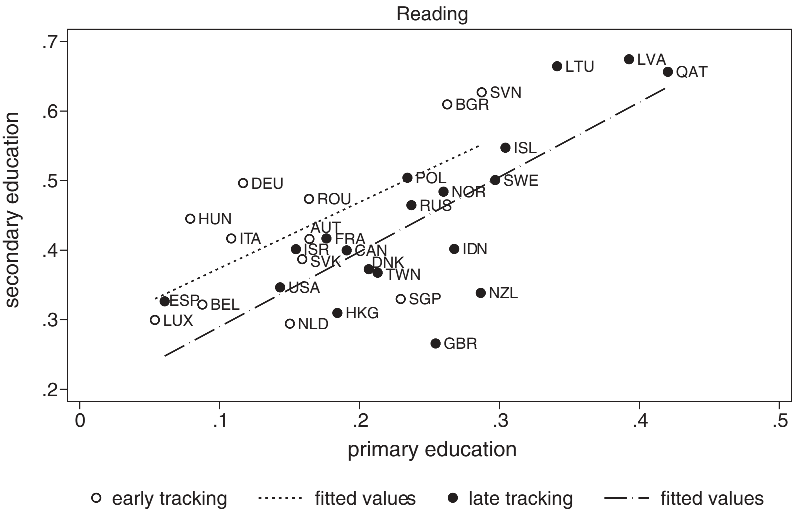

Figure 1. The gender test score gap (F-M) in reading in primary and secondary education, and early tracking.

These results stand in sharp contrast to the patterns of the baseline model, as early tracking is significantly related to the gender slope of test scores.

The key variable here is the triple interaction term of tracking, secondary level education and female student. Its coefficient is statistically significant for each subject in both specifications. This indicates that in tracking countries the gender gap evolves in a way significantly different to that in the non-tracking group from primary to secondary education.

The triple interaction term has a negative effect, suggesting that later tracking impairs the performance of girls relative to boys. In mathematics, for example, an increase in the female-male difference in test scores from grade 4 to age 15 is 0.0663 SD smaller in non-tracking countries than in early-tracking ones. The dummy variable specifications tell the same story: in non-tracking countries, girls' advantage in reading decreases, while the gap in mathematics widens.

Overall, these results suggest that girls gain with early tracking relative to boys. This is not surprising, as boys enrol in vocational tracks more often than girls. Consequently, after tracking more boys than girls receive a lower level and lower quality of schooling in academic subjects.

It is important to emphasise that these effects represent the direct causal impact of tracking. The multilevel model also allows us to estimate the general association of tracking and the gender gap, net of this direct effect, at the same time. The coefficients of the double interaction terms in Table 4 suggest that in early tracking countries girls tend to perform relatively worse than boys in reading and science before tracking takes place. For mathematics, the coefficients are not significant, but are similar in magnitude, with the same sign. These effects can hardly be attributed to tracking itself. Instead, they imply that some other features of the education system, correlated with early tracking, generate relative advantages for boys in these countries.

Here, it should be noted that the direct effect of tracking and the effect of its unobserved correlates have opposite signs. In the baseline model, the sum of these two effects was estimated, and they were found to cancel out, resulting in no relationship at age 15.

In summary, the implication is that in early tracking countries boys' relative advantage over girls is larger in primary school compared to non-tracking countries, but later boys suffer losses due to tracking. As these two effects offset each other, there is no correlation at age 15.

It is also interesting to compare the coefficients of the other two policy variables with those estimated in the baseline model. These variables have the same interpretation, the two models differ only in the sample. The effects for a restricted set of countries are estimated here, while the sample contains two age cohorts of students. In spite of these differences, the results are very similar.

4.4. Individualised teaching

Finally, we turn to individualised teaching. The baseline model shows that in a cross-country comparison more student-oriented teaching practices seem to benefit girls in each of the three subjects. In contrast to grade retention and early tracking, there is no straightforward way to provide further evidence concerning these factors at the country level. Hence we are looking at individualised teaching effects within countries. It is

Table 4. Difference-in-differences estimates of the effect of early tracking on the gender test score gap.

	Maths		Reading		Science	
	(1)	(2)	(3)	(4)	(5)	(6)
female X log grade retention	−0.0323*	−0.0330*	−0.0324*	−0.0329*	−0.0217	−0.0233
	(0.0177)	(0.0172)	(0.0177)	(0.0184)	(0.0190)	(0.0190)
female X tracking age	0.00860		0.0113*		0.0213***	
	(0.00744)		(0.00595)		(0.00666)	
female X non-tracking		0.0342		0.0475**		0.0756***
		(0.0284)		(0.0241)		(0.0293)
female X individualised teaching	0.125***	0.128***	0.106**	0.106**	0.120**	0.123**
	(0.0405)	(0.0421)	(0.0525)	(0.0482)	(0.0488)	(0.0487)
female X tracking age X PISA	−0.0137*		−0.0156*		−0.0203**	
	(0.00724)		(0.00899)		(0.00860)	
female X non-tracking X PISA		−0.0663**		−0.0679**		−0.0889***
		(0.0297)		(0.0311)		(0.0329)
Observations	350,562	350,562	396,189	396,189	350,562	350,562
Number of countries	27	27	30	30	27	27

Country-level variables as in Table 2. Additional controls: indicator variable of PISA observations, female students, and the interaction of PISA observations and female students. Robust standard errors clustered at the country-level are given in parentheses.

***p < 0.01, **p < 0.05, *p < 0.1

assumed that if this factor and related policies do indeed affect the gender gap, the effect can be recognised at the school and student level too, since in most countries there is ample variation in individualised teaching both between and within schools. However, due to potential selectivity and endogeneity biases these estimates should not be interpreted as evidence of a causal relationship.

In order to estimate the effect within countries, the baseline model was extended by the addition of a third level, that of schools (see Appendix 3 for the detailed model specification).

In this approach, within country and between school variance in teaching practices is exploited. The within country effect is represented by the interaction between gender and student-oriented teaching at the school level.

A major problem with this approach is that neither students nor teachers can be expected to be randomly distributed across schools. Teachers are often matched to students in a non-random fashion, and the sorting of students and teachers results in selection bias in the estimation of the effects of teaching practices and school characteristics (Kane et al. 2011). To mitigate these biases a second model was analysed, relying on within-school variation only, which is independent of sorting across schools. As classes cannot be identified in PISA, we rely on student-level variation here, which presumably mainly reflects differences between classes. In this second specification, an index of student-oriented teaching and its interaction with gender at the student level is added. The coefficient of this interaction term represents the within-school effect.

These models were estimated for mathematics scores only, as in PISA 2012 teacher behaviour was measured for mathematics lessons. While at the country level these variables are likely to be appropriate proxies for teacher behaviour in general, this is less likely the case within countries, at the school or class level. For example, a mathematics teacher in class A employing more student-oriented practices than the mathematics teacher in class B is probably a very weak predictor of the difference in the behaviour of the science teachers in the two classes. Hence we confine the within-country analysis to mathematics.

Table 5 shows the results. In the within-country models student-oriented teaching in the school has a significant impact on the gender slope (Column 1 of Table 5). The more prevalent individualised teaching practices are, the better girls perform in mathematics relative to boys. A unit increase of the student-oriented teaching index implies an increase of 0.0551SD in the female-male test score difference. At the same time, student-oriented teaching practices go together with a lower overall level of test scores.

The within-school effects reflect the same pattern (Column 2 of Table 5). Girls seem to benefit more from individualised teaching relative to boys. In these models, the school mean of student-oriented teaching is not significantly related to the gender slope due to multicollinearity; the student- and school level measures are highly correlated.

Altogether, within-country and within-school estimates are in line with the country-level effects estimated in the baseline model. More student-oriented teaching practices appear to improve the test scores of girls relative to boys significantly. Though causal effects cannot be identified here, this evidence lends further support to the supposition that more student-oriented teaching practices are indeed relatively beneficial for girls and reduce the test score gap in mathematics.

5. Conclusions

Based on the 2012 wave of PISA data, the relationships between different educational policies and the gender gap in test scores were assessed from a cross-country perspective. The analysis covered all three fields of competence measured in PISA: mathematics, reading and science. The effects of three educational policies that education systems use to manage student heterogeneity were examined: early tracking, grade retention and individualised teaching.

In this study, a two-stage empirical strategy was pursued. First, the association between the three policy variables and the gender gap was analysed using a simple multilevel model. Further evidence on the impact of each policy variable was then examined by extending the model in different ways.

Our results can be summarised in four key findings. First, at the country level grade retention is closely related to the gender test score gap. The widespread use of grade retention appears to favour boys; in countries with high grade retention rates boys perform better than girls in all three subjects. However, further evidence indicates that this correlation is very unlikely to represent a causal effect. Second, although we found no association between tracking and the gender test score gap at the age of 15, boys tend to perform relatively better in early tracking countries at grade 4, long before tracking occurs. This finding is in accordance with Bedard and Cho (2010). Overall, neither of these correlations represent a causal effect. These effects may be mediated by other factors omitted from the analysis (i.e. factors that are correlated with grade retention or early tracking on the one hand, and gender gap on the other).

In a broader context these two findings suggest that more selective (Mons 2007), or in OECD (2013b) terminology, more stratified school systems tend to favour boys due to factors outside of the direct effects of certain educational policies. In other words, it is not stratification policies themselves that increases the male-female gender gaps, but other factors associated with them.

Table 5. Within-country and within school effects of student-oriented teaching practices on the gender gap in mathematics test scores.

	Within-country model	Within-school model
	(1)	(2)
female	−1.170***	−1.109***
	(0.214)	(0.209)
student-oriented teaching		−0.176***
		(0.00899)
female X student-oriented teaching		0.0370***
		(0.00735)
student-oriented teaching (school mean)	−0.385***	−0.241***
	(0.0244)	(0.0280)
female X student-oriented teaching (school mean)	0.0551***	0.0189
	(0.0135)	(0.0183)
Observations	470,944	306,279
Number of schools	17,901	17,901
Number of countries	62	62

The models include student-level controls and country-level variables and interactions as in Table 2. School-level controls are mean ESCS, the share of girls, private school status, urban location and the share of students at the upper-secondary level, and interactions with female student. Robust standard errors clustered at the country-level are shown in parentheses.
***p < 0.01, **p < 0.05, *p < 0.1

What might be these factors? First, it is possible that in countries with a stratified educational system parental and teacher attitudes and performance expectations convey gender norms in a more compelling way, generating larger gender differences in student aspiration and motivation.

Another explanation may point to school climate, teachers' attitudes and teaching practices. In early tracking regimes or education systems with a high rate of grade retention a more orderly school climate may benefit boys relative to girls. Duckworth et al. (2015) found that boys' lower self-control contributes to female advantage in school performance. A more disciplined school climate may to some extent substitute for motivation and self-control in case of boys.

Out third key finding on teaching practices indirectly confirms this interpretation. We found that more individualised teaching practices tend to improve the performance of girls relative to boys. This association can be observed both at the country level and within countries. Though a causal effect cannot be identified here, a direct impact is likely to exist, given the suggestive evidence from within-country and within-school models.

Individualised teaching practices are more prevalent in non-stratified educational systems, as these countries rely on these practices to cope with student heterogeneity instead of using different forms of selection (Mons 2004, 2007; Janmaat and Mons 2011). Our results show that these practices are associated with a larger female advantage both across and within countries – a mirror image of boys performing better in stratified systems. Consequently, the lower incidence of individualised teaching practices in selective school systems may be part of the missing link between stratification and boys' relative advantage in school.

Our fourth key finding is that early tracking has a marked direct effect on gender test score gap. Analysing the evolution of the gender gap from primary to secondary education provides strong evidence for early tracking directly benefiting girls relative to boys in each subject. This means that early selection tends to increase the gender gap in reading, while narrowing the gender gaps in mathematics and science.

At first sight this result appears to contradict to the conclusion of previous works. Van Hek, Buchmann, and Kraaykamp (2019) claim that early tracking benefit boys in reading at the age of 15. Bedard and Cho (2010) report a similar association for mathematics and science in grade 8. However, these studies estimate the overall correlation between tracking age and gender gap, while we identified the direct effect of tracking, representing a causal relationship. The overall correlation reflects the sum of this direct effect and the correlation between stratification and the gender gap (see above), explaining the difference in the results.

Another strand of literature demonstrates a positive association between late tracking and female educational advantage across countries (Hadjar and Buchmann 2016) as well as within countries (Pekkarinen 2008; Scheeren, Van de Werfhorst, and Bol 2018). Pekkarinen (2008) analysed the effect of a comprehensive school reform in Finland, while Scheeren, Van de Werfhorst, and Bol (2018) investigated similar reforms in several European countries. These two studies employ a credible identification strategy to estimate a causal effect of tracking age. Though our study used a different outcome variable, the different results leads to the question of how early tracking can directly benefit girls in test scores and increase boys' relative advantage in educational attainment at the same time?

First, gender norms may overwrite the effect of school performance in educational choices after secondary school, generating relative advantages for boys in educational attainment. As early tracking can be expected to reinforce the effect of gender norms (Scheeren, Van de Werfhorst, and Bol 2018), this may offset the relative gains for girls in terms of test scores.

Second, the direct effect of tracking might well depend on the institutional and social context. We analysed the gender test score gap in a cohort that entered upper-secondary education in the early 2010s. In this period girls tend to outnumber boys at the academic tracks in most of the countries (see also Pekkarinen 2008 for Western European countries; Jürges and Schneider 2011 for Germany), while they are heavily underrepresented at the vocational tracks.[5] The direct effect of tracking age on the gender test score gap is likely to emerge from this inequality in track placement. If the academic track provides education of higher quality, and girls are overrepresented at this track, girls benefit more from tracking.

Meanwhile, the school reforms in Finland and in most of the countries analysed by Scheeren, Van de Werfhorst, and Bol (2018) took place several decades earlier, in the 1970s and 1980s. In that period sorting across tracks was less biased towards girls. In Germany, for example, the share of girls on the academic track has increased constantly since 1940, but until the 1970s boys were overrepresented (Jürges, Reinhold, and Salm 2011). Moreover, in that period gender norms could have a strong effect on track choice, resulting in early tracking amplifying gender differences in educational attainment (Scheeren, Van de Werfhorst, and Bol 2018).

The two parallel trends of the increased share of girls on the academic track, and the decreased effect of gender norms on educational choices may explain at least part of the difference between our result and the findings of the education reform studies. In line with this argument, van Elk, van der Steeg, and Webbink (2011) found no effect of tracking on the gender gap in completion of higher education in the 1990s in the Netherlands.

Altogether, our results confirm that educational institutions and policies do matter in shaping gender inequalities in school performance. At the same time, the findings call for further research on how features of the educational system affect the gender gap. First, more research is needed to gain a better understanding of the interplay among the broader educational policy and social contexts and early tracking. Second, though the present study focused on tracking that takes place between schools, further research should look at other forms of tracking (i.e. ability grouping within schools). Third, there is little knowledge on the effects of teaching practices on the gender gap in test scores. Our finding is limited to the gender gap in mathematics, and it is not clear whether the same holds true for other subjects. Forth, future research should look at the mechanisms through which teaching practices influence student achievement. Understanding them is would be very valuable for education policy, as improving teaching practices through teacher training based on this evidence provides an opportunity to foster gender equality in schools.

Notes

1. Here tracking refers to streaming students into different schools with either academic or vocational focus.
2. Using unstandardised test scores to calculate the gender gap leads to similar estimation results (results available upon request).
3. These results seem to contradict the findings of Van Hek, Buchmann, and Kraaykamp (2019)), who reports a positive effect of the tracking age on the gender slope in reading. However, she estimated this positive effect in a three-level model including schools as a separate level and, thus, controlling for sorting across schools. In that setting, the positive effect is conditional on sorting. In contrast, the two-level model here represents the unconditional association. It is to be noted that schools play an important mediating role, as sorting is part of the mechanism behind the tracking effect (Skopek and Dronkers 2015). Hence, in order to estimate the total effect, sorting across schools should not be controlled for.
4. An indicator for non-tracking is used instead of early tracking to have a coefficient with similar sign to tracking age.
5. In the 2012 PISA sample of early tracking countries the average share of girls on the academic track is 52 percent, as opposed to 39 and 43 percent on pre-vocational and vocational tracks. The differences are even higher in European countries that give the majority of the subsamples we used to estimate the direct effect of tracking.

Acknowledgments

We thank Hedvig Horváth, Dániel Horn, Péter Róbert and Júlia Varga for their helpful comments on earlier versions of the paper. Also, we are grateful for the constructive suggestions of the anonymous reviewer.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Hungarian National Research Development and Innovation Fund [grant number K109338] and the European Union's Horizon 2020 research and innovation programme [grant number 691676 (EdEN)].

Notes on contributors

Zoltán Hermann is a research fellow at the Institute of Economics of Centre for Economic and Regional Studies (IE/CERS) and an associate professor at the Centre for Labour Economics of Corvinus University of Budapest. His main research interests are economics of education and local public economics, especially the production of human capital, inequalities in education and educational institutions, policy evaluation and financing education.

Marianna Kopasz is a research fellow at the Institute for Political Science, Centre for Social Sciences of the Hungarian Academy of Sciences. She holds a Ph.D. in Sociology from the Corvinus University of Budapest. Her main research interests comprise child welfare, social policy, and educational policy.

ORCID

Zoltán Hermann http://orcid.org/0000-0001-9805-9905
Marianna Kopasz http://orcid.org/0000-0001-8661-5644

References

Ammermüller, A. 2005. "Educational Opportunities and the Role of Institutions." *ZEW Discussion Papers 05–44*.

Autor, D., D. Figlio, K. Karbownik, J. Roth, and M. Wasserman. 2016. "School Quality and the Gender Gap in Educational Achievement." *American Economic Review* 106 (5): 289–295. doi:10.1257/aer.p20161074.

Ayalon, H., and I. Livneh. 2013. "Educational Standardization and Gender Differences in Mathematics Achievement: A Comparative Study." *Social Science Research* 42 (2): 432–445. doi:10.1016/j.ssresearch.2012.10.001.

Baye, A., and C. Monseur. 2016. "Gender Differences in Variability and Extreme Scores in an International Context." *Large-scale Assessments in Education* 4: 1. doi:10.1186/s40536-015-0015-x.

Bedard, K., and I. Cho. 2010. "Early Gender Test Score Gaps across OECD Countries." *Economics of Education Review* 29 (3): 348–363. doi:10.1016/j.econedurev.2009.10.015.

Bol, T., and H. G. van de Werfhorst. 2013. "Educational Systems and the Trade-off between Labor Market Allocation and Equality of Educational Opportunity." *Comparative Education Review* 57 (2): 285–308. doi:10.1086/669122.

Bryan, M. L., and S. P. Jenkins. 2015. "Multilevel Modelling of Country Effects: A Cautionary Tale." *European Sociological Review* 32 (1): 3–22. doi:10.1093/esr/jcv059.

Castejón, A., and A. Zancajo. 2015. "Educational Differentiation Policies and the Performance of Disadvantaged Students across OECD Countries." *European Educational Research Journal* 14 (3–4): 222–239. doi:10.1177/1474904115592489.

Duckworth, A. L., E. P. Shulman, A. J. Mastronarde, S. D. Patrick, J. Zhang, and J. Druckman. 2015. "Will Not Want: Self-control Rather than Motivation Explains the Female Advantage in Report Card Grades." *Learning and Individual Differences* 39: 13–23. doi:10.1016/j.lindif.2015.02.006.

Dupriez, V., X. Dumay, and A. Vause. 2008. "How Do School Systems Manage Pupils' Heterogeneity?" *Comparative Education Review* 52 (2): 245–273. doi:10.1086/528764.

Else-Quest, N. M., J. S. Hyde, and M. C. Linn. 2010. "Cross-national Patterns of Gender Differences in Mathematics: A Meta-analysis." *Psychological Bulletin* 136 (2): 102–127. doi:10.1037/a0018053.

Fryer, R., and S. Levitt. 2010. "An Empirical Analysis of the Gender Gap in Mathematics." *American Economics Journal: Applied Economics* 2 (2): 210–240.

González de San Román, A., and S. De la Rica Goiricelaya. 2012. "Gender Gaps in PISA Test Scores: The Impact of Social Norms and the Mother's Transmission of Role Attitudes." *IZA Discussion Paper No. 6338*.

Guiso, L. F., P. Monte, P. Sapienza, and L. Zingales. 2008. “Culture, Gender, and Math.” *Science* 320 (5880): 1164–1165. doi:10.1126/science.1154094.

Hadjar, A., and C. Buchmann. 2016. “Education Systems and Gender Inequalities in Educational Attainment.” In *Education Systems and Inequalities: International Comparisons*, edited by A. Hadjar and C. Gross, 159–184. Bristol: Policy Press. doi:10.2307/j.ctt1t892m0.

Hanushek, E. A., and L. Woessmann. 2006. “Does Educational Tracking Affect Performance and Inequality? Differences-in-differences Evidence across Countries.” *Economic Journal* 116 (510): C63–C76. doi:10.1111/j.1468-0297.2006.01076.x.

Hanushek, E. A., and L. Woessmann. 2010. “The Economics of International Differences in Educational Achievement.” In *Handbook of the Economics of Education*, edited by E. A. Hanushek, S. J. Machin, and L. Woessmann, 89–199. Vol. 3. North Holland: Elsevier.

Hidalgo-Cabrillana, A., and C. Lopez-Mayan. 2015. “Teaching Styles and Achievement: Student and Teacher Perspectives.” *Working Paper 2/15*. Universidad Autónoma de Madrid, Spain.

Hochweber, J., and S. Vieluf. 2018. “Gender Differences in Reading Achievement and Enjoyment of Reading: The Role of Perceived Teaching Quality.” *The Journal of Educational Research* 111 (3): 268–283. doi:10.1080/00220671.2016.1253536.

Huang, M.. 2009. “Classroom homogeneity and the distribution of student math performance: a country-level fixed-effects analysis.” *Social Science Research*, 38 (4): 781–791. https://doi.org/10.1016/j.ssresearch.2009.05.001

Ikeda, M., and E. García. 2014. “Grade Repetition: A Comparative Study of Academic and Non-academic Consequences.” *OECD Journal: Economic Studies* 2013/1: 269–315. doi:10.1787/eco_studies-2013-5k3w65mx3hnx.

Janmaat, J. G., and N. Mons. 2011. “Promoting Ethnic Tolerance and Patriotism: The Role of Education System Characteristics.” *Comparative Education Review* 55 (1): 56–81. doi:10.1086/657105.

Jimerson, S. R., G. E. Anderson, and A. D. Whipple. 2002. “Winning the Battle and Losing the War: Examining the Relation between Grade Retention and Dropping Out of High School.” *Psychology in the Schools* 39 (4): 441–457. doi:10.1002/pits.10046.

Jimerson, S. R., S. M. W. Pletcher, K. Graydon, B. L. Schnurr, A. B. Nickerson, and D. K. Kundert. 2006. “Beyond Grade Retention and Social Promotion: Promoting the Social and Academic Competence of Students.” *Psychology in the Schools* 43 (1): 85–97. doi:10.1002/pits.20132.

Jürges, H., S. Reinhold, and M. Salm. 2011. “Does Schooling Affect Health Behavior? Evidence from the Educational Expansion in Western Germany.” *Economics of Education Review* 30 (5): 862–872. doi:10.1016/j.econedurev.2011.04.002.

Jürges, H., and K. Schneider. 2011. “Why Young Boys Stumble: Early Tracking, Age and Gender Bias in the German School System.” *German Economic Review* 12 (4): 371–394. doi:10.1111/j.1468-0475.2011.00533.x.

Kane, T. J., E. S. Taylor, J. H. Tyler, and A. L. Wooten. 2011. “Identifying Effective Classroom Practices Using Student Achievement Data.” *Journal of Human Resources* 46 (3): 587–613. doi:10.1353/jhr.2011.0010.

Korbel, V., and M. Paulus. 2017. “Do Teaching Practices Impact Socio-emotional Skills?” *IES Working Paper 4/2017. IES FSV*. Charles University.

Lavrijsen, J., and I. Nicaise. 2015. “New Empirical Evidence on the Effect of Educational Tracking on Social Inequalities in Reading Achievement.” *European Educational Research Journal* 14 (3–4): 206–221. doi:10.1177/1474904115589039.

Manacorda, M. 2012. “The Cost of Grade Retention.” *Review of Economics and Statistics* 94 (2): 596–606. doi:10.1162/REST_a_00165.

Marks, G. N. 2008. “Accounting for the Gender Gaps in Student Performance in Reading and Mathematics: Evidence from 31 Countries.” *Oxford Review of Education* 34 (1): 89–109. doi:10.1080/03054980701565279.

Martin, A. J. 2009. “Age Appropriateness and Motivation, Engagement, and Performance in High School: Effects of Age within Cohort, Grade Retention, and Delayed School Entry.” *Journal of Educational Psychology* 101 (1): 101–114. doi:10.1037/a0013100.

Mons, N. 2004. “De l’école unifiée aux écoles plurielles: évaluation internationale des politiques de différenciation et de diversification de l’offre éducative.” Doctoral thesis, Université de Bourgogne.

Mons, N. 2007. *Les nouvelles politiques éducatives: La France fait-elle les bons choix?* Paris: Presses Universitaires de France.

Morrison, K., and A. I. O. No. 2007. “Does Repeating a Year Improve Performance? the Case of Teaching English.” *Educational Studies* 33: 353–371. doi:10.1080/03055690701423333.

OECD. 2013a. *PISA 2012 Results: Ready to Learn: Students’ Engagement, Drive and Self-Beliefs*. Vol. III vols. Paris: OECD Publishing. doi:10.1787/19963777.

OECD. 2013b. *PISA 2012 Results: What Makes Schools Successful? Resources, Policies and Practices*. Vol. IV vols. Paris: PISA, OECD Publishing. doi:10.1787/19963777.

OECD. 2015. *The ABC of Gender Equality in Education: Aptitude, Behavior, Confidence*. Paris: OECD, PISA, OECD Publishing. doi:10.1787/9789264229945-en.

Pekkarinen, T. 2008. “Gender Differences in Educational Attainment: Evidence on the Role of Tracking from a Finnish Quasi-experiment.” *Scandinavian Journal of Economics* 110 (4): 807–825. doi:10.1111/j.1467-9442.2008.00562.x.

Pekkarinen, T. 2012. “Gender Differences in Education.” *IZA Discussion Paper No. 6390*.

Penner, A. M. 2008. “Gender Differences in Extreme Mathematical Achievement: An International Perspective of Biological and Social Forces.” *American Journal of Sociology* 114 (S1): S138–S170. doi:10.1086/589252.

Riegle-Crumb, C. 2005. “The Cross-national Context of the Gender Gap in Math and Science.” In *The Social Organization of Schooling*, edited by L. V. Hodges and B. Schneider, 227–243. New York: Russell Sage Foundation.

Scheeren, L., H. G. Van de Werfhorst, and T. Bol. 2018. “The Gender Revolution in Context: How Later Tracking in Education Benefits Girls.” *Social Forces* 97 (1): 193–220. doi:10.1093/sf/soy025.

Schnepf, S. V. 2004. “Gender Equality in Educational Achievement: An East-West Comparison.” *IZA Discussion Paper No. 1317*.

Schütz, G., H. W. Ursprung, and L. Woessmann. 2008. “Education Policy and Equality of Opportunity.” *Kyklos* 61 (2): 279–308. doi:10.1111/j.1467-6435.2008.00402.x.

Schwerdt, G., and A. C. Wuppermann. 2011. “Is Traditional Teaching Really All that Bad? A Within-student Between-subject Approach.” *Economics of Education Review* 30 (2): 365–379. doi:10.1016/j.econedurev.2010.11.005.

Skopek, J., and J. Dronkers. 2015. “Performance in Secondary School in German States: A Longitudinal Three-Level Model.” *Working Paper*, August. doi:10.5157/neps:sc3:2.0.0.

Stoet, G., and D. C. Geary. 2015. “Sex Differences in Academic Achievement are Not Related to Political, Economic, or Social Equality.” *Intelligence* 48: 137–151. doi:10.1016/j.intell.2014.11.006.

Van de Werfhorst, H. G. 2015. “Institutional Contexts for Socio-Economic Effects on Schooling Outcomes.” In *Emerging Trends in the Social and Behavioral Sciences: An Interdisciplinary, Searchable, and Linkable Resource*, edited by R. Scott and S. Kosslyn. New York: John Wiley & Sons.

van Elk, R., M. van der Steeg, and D. Webbink. 2011. “Does the Timing of Tracking Affect Higher Education Completion?” *Economics of Education Review* 30 (5): 1009–1021. doi:10.1016/j.econedurev.2011.04.014.

Van Hek, M., C. Buchmann, and G. Kraaykamp. 2019. “Educational Systems and Gender Differences in Reading: A Comparative Multilevel Analysis.” *European Sociological Review* 35 (2): 169–186. doi:10.1093/esr/jcy054.

Van Langen, A., R. Bosker, and H. Dekkers. 2006. “Exploring Cross-national Differences in Gender Gaps in Education.” *Educational Research and Evaluation* 12 (2): 155–177. doi:10.1080/13803610600587016.

Waldinger, F. 2007. *Does Ability Tracking Exacerbate the Role of Family Background for Students Test Scores?* Mimeo, London School of Economics.