Essays on the Economics of Education

David Contreras Gomez

A thesis submitted for the degree of Doctor of Philosophy

School of Economics
University of Nottingham

December 2021

Declaration

I certify that the work submitted is solely my own work other than where I have clearly indicated that it is the work of others.

The copyright of this thesis rests with the author. Quotation from it is permitted, provided that full acknowledgement is made.

Statement of co-authored work:

I confirm that Chapter 2 was jointly co-authored with Professor Facundo Albornoz and Professor Richard Upward.

Acknowledgements

I would like to thank my supervisors, Professor Facundo Albornoz and Professor Richard Upward, for their invaluable guidance and feedback on the course of my research and for their support and trust throughout this process.

I would also like to express my sincere gratitude to the Chilean government and the National Agency for Research and Development (ANID) for providing me with the financial support that made this project possible.

I am also grateful to the Research Centre of the Ministry of Education, to the Education Quality Agency of Chile, to the Department of Assessment, Evaluation and Educational Records (DEMRE) of the University of Chile for granting access to the data used in this thesis.

Thanks to my colleagues in the School of Economics for the supportive environment they offered me. I am particularly thankful for the administrative support provided by Hayley McCalla and Natalia Bas.

Finally, a special thanks to my family for their confidence and continuous encouragement. My deepest appreciation goes to my wife, Pamela, for his love and endless support. Thank you for believing in me and for all those fruitful discussions that undoubtedly improved this thesis.

Abstract

This dissertation studies different topics in education policy, using precise and comprehensive administrative data from Chile. The thesis contains three chapters.

Chapter 1 examines the presence of systematic differences in teachers' grading behaviour across gender and whether these can be attributed to teacher bias. This chapter measures these differences by comparing teachers' grades with national exams, which are externally and anonymously marked. Consistent with the literature, the gender gap in teacher grading is against boys. Using a dataset with gender gaps at class-subject level – which allows to follow teachers in different classes over time – the chapter shows that teachers' grading behaviour is far from being a fixed characteristic of the teachers. Instead, grading gaps can be attributed either by teachers rewarding differently students' behaviour based on their gender, or by female students being more effective in transforming these behaviours into higher school grades.

Chapter 2 explores the effectiveness of repeating the student-teacher match on students' test scores. The analysis uses detailed information on all student-teacher matches across multiple subjects and multiple years, and a national anonymous measure of student test scores which is uncontaminated by any teacher or school biases in grading. This chapter exploits a plausibly exogenous source of variation in the process of matching teachers to students which arises because of a discontinuity in teacher retention at the legal retirement age, and finds that repeating the student-teacher match has a robust positive effect on test scores which aggregates up to the student, class, and school-level. Also, repeating the match also has a positive effect on attendance, student behaviour and teacher expectations.

Chapter 3 focuses on the role of family in the education of children. In particular, this chapter explores how children affect the long-term educational outcomes of their siblings. To identify these effects, this chapter exploits an exogenous variation in children's school starting age caused by Chile's school entry cutoff date. The chapter shows that younger siblings have a large and significant impact on older siblings scores in college admission exams, while no spillover effect is found from older-to-younger siblings. Finally, it discusses potential channels, emphasising the role of parental investments and the timing of school start in explaining the results.

Contents

List of Tables

List of Figures

Abbreviations

CLP	Chilean Peso
DD	Difference-in-Differences
DEMRE	*Departamento de Medición, Registro y Evaluación* Department of Assessment, Evaluation and Educational Records
FE	Fixed Effects
GDP	Gross Domestic Product
GPA	Grade Point Average
IRT	Item Response Theory
ITT	Intention-to-treat
LRA	Legal Retirement Age
MDS	*Ministerio de Desarrollo Social* Ministry of Social Development
MSE	Mean Squared Error
OECD	Organisation for Economic Co-operation and Development
PIRLS	Progress in International Reading Literacy Study
PISA	OECD Programme for International Student Assessment
PSU	*Prueba de Selección Universitaria* University Selection Test
RD	Regression Discontinuity
SES	Socioeconomic Status
SSA	School Starting Age
SIMCE	*Sistema de Medición de la Calidad de la Educación* System to Measure the Quality of Education
TERCE	*Tercer Estudio Regional Comparativo y Explicativo* Third Regional Comparative and Explanatory Study
TIMSS	Trends in International Mathematics and Science Study
UNESCO	United Nations Educational, Scientific and Cultural Organization
US	United States

Introduction

The role of human capital and knowledge on economic development has long been the subject of attention of economists. A wide body of literature has recognized the accumulation of human capital as the main source of economic growth, as well as the one responsible for technological change and improvements in productivity (e.g. Goldin and Katz, 2008; Hanushek and Woessmann, 2008; Acemoglu and Autor, 2012). In this context, education systems focused on the acquisition of skills and abilities are certainly one of the cornerstones of human capital. Also, investments in education and human capital play an important equalizing role, reducing and tackling the sources of inequality. As such, previous studies have established the importance of education in addressing income segregation and promoting intergenerational mobility (e.g. OECD, 2018; Breen, 2019; Chetty et al., 2020).

The following chapters focus their analysis on the Chilean educational system. Chile is a middle-income country with the highest GDP per capita after adjusting for purchasing power parity in Latin America (USD $25,067$ in 2020) (World Bank, 2021). Over the past decades Chile has experienced significant improvements in its quality of life, supported by a consistent reduction in poverty (it fell from 68.5% in 1990 to 8.6% in 2017), macroeconomic stability and structural reforms in the social sector (Schmidt-Hebbel, 2006; Ffrench-Davis, 2010; MDS, 2020). Despite these improvements, and like most countries in Latin America, income inequality in Chile remains high. Measured by the Gini index after taxes, income inequality in Chile is one of the highest among the OECD countries (OECD, 2018).

In recent years, Chile's education system has succeeded in expanding coverage and improving performance, but income-related gaps in educational attainments remain one of the main challenges. The country has increased access to education and school enrolment across all levels of education: primary and secondary education are almost universal and tertiary education – although below the OECD average – has increased significantly in the last two decades (MDS, 2017; OECD, 2017). This has been accompanied by improvements in student achievement in international tests, as shown the results of the 2013 UNESCO-TERCE study (UNESCO, 2015) and the PISA tests (OECD, 2019*a*). In these tests, Chile scores above the regional average but remains at the bottom end within the OECD countries. However, inequalities are also present in the Chilean education system. Several studies have suggested that Chile presents high levels of socio-economic sorting across schools (e.g. Hsieh and Urquiola, 2006; Bellei et al., 2010; Mizala and Torche, 2012). Data from PISA 2018 reflect these educational inequities by showing a stronger relationship between socio-economic background and performance compared to other OECD countries (OECD, 2019*b*). Educational resources are also unequally distributed across school types (e.g. public and private-subsidised schools) and geographical location (e.g. rural and urban schools). In particular, prior research has suggested the existence of negative selection of teachers into socio-economically disadvantaged schools (e.g. Correa et al., 2015; Behrman et al., 2016). Disadvantaged schools also have higher teacher shortages, and are more likely to present inadequate infrastructure or lack of sufficient teaching materials (Santiago et al., 2017). Despite the efforts undertaken by successive administrations and the ongoing reforms, the education system in Chile faces challenges on multiple fronts, including strengthening the quality and coverage of early childhood education, improving teacher professional development, and improving equity in access to high-quality post-secondary education (OECD, 2017).

Considering all the above, this research focuses on studying different topics in education policy, with the aim of contributing to the public policy debate on education. To this end, this thesis takes advantage of unique, precise, and comprehensive administrative data in the Chilean context, in which students, teachers and schools can be followed longitudinally. This research benefits from these data, and presents three well-identified empirical studies designed to enhance our understanding of education in developing countries as Chile. In what follows, the aims and main contributions of the three empirical studies are summarised.

The first chapter adds to the existing literature that uses girls' and boys' gaps in non-blind and blind assessments to measure gender-stereotypical behaviour. This chapter uses detailed student-level administrative data from Chile to address two major questions.

First, it investigates whether teachers' grading behaviour systematically differs across gender. It shows that boys get lower grades than girls when they are assessed by their teachers compared to their scores in the national exams. This result is in line with previous studies in the area (e.g. Lavy, 2008; Falch and Naper, 2013; Terrier, 2020). Additionally, it documents that the gender grading gap against boys is robust across subjects, school type, geographical location, and school size. Finally, when examining how the gender gap in grading changes with the age of students, it indicates that these gaps present a clear pattern, increasing with age up to the end of primary school and decreasing in high school.

Second, it discusses whether gender grading gaps are capturing teachers' biased behaviour, or they are reflecting differences in students' behaviour. Many researchers have utilised systematic differences between non-blind and blind assessments across groups (e.g. gender) as a measure of teachers' stereotypes (e.g. Lavy, 2008; Burgess and Greaves, 2013; Botelho et al., 2015; Terrier, 2020). This study challenges this idea, by theoretically showing that deviations between school grades (i.e. non-blind tests) and standardised tests (i.e. blind tests) might reflect not only teacher's biases, but also differences in how these two types of evaluations are linked to factors related to students' achievement, such as student effort, ability, and behaviour.

The hypothesis that gender gaps in grading are capturing teachers' biased behaviour is tested empirically by taking advantage of the fact that teachers are observed in multiple classes over several years. The results show that teachers' grading behaviour is not persistent across classes. This suggests that teachers' grading behaviour is not a fixed characteristic of the teacher. In turn, it documents that differences in gender grading gaps are systematically related to the characteristics of the class.

This chapter argues that gender gaps in grading are explained by differences in students' behaviour. Specifically, it shows that part of the gender grading gap is due to differential effort – i.e. girls exert more effort than boys – and part is due to girls being rewarded more for a given amount of effort – i.e. even if boys put in the same effort as girls, they would not get the same school grade.

The findings of this study have significant policy implications. Firstly, gender gaps in grading compromise the extent to which school grades inform about children's skills and learning. Secondly, in Chile – as in many other countries – school grades form part of the criteria for admission to higher education, and therefore have an impact on the likelihood of attendance. Given this, the evidence of gender gaps in grading encourages a discussion on the necessity of revising schools grading procedures and policies.

The second chapter evaluates the effects of "looping" on student achievement. Looping is an instructional design in which a teacher stays with the same group of students for two or more years. Although there is a very strong consensus in the educational literature that looping improves student outcomes (e.g. Nichols and Nichols, 2002; Cistone and Shneyderman, 2004; Tucker, 2006; Franz et al., 2010), much of the research up to now has been descriptive. The analysis in this chapter, therefore, fills this gap by providing a systematic assessment of the causal effect of looping on student achievement. Estimating the causal effect of repeating the student-teacher match is challenging because: (i) student-teacher matches are non-randomly selected, and (ii) new matches are selected from a pool which includes teachers who are new in the school.

This chapter addresses these concerns by exploiting a unique and rich student-teacher data for 8th graders in Chile, using two empirical approaches. First, it uses a multidimensional fixed-effects model which controls for selection of students and teachers into repeat matches, however, it does not fully mitigate the concern that school managers (or teachers) might decide to repeat matches based on the quality of the match. Second, and to solve this problem, this chapter uses a regression discontinuity design based on pension eligibility rules, that changes the probability of repeating the student-teacher match. In particular, grade 8 students whose grade 7 teacher reached the legal retirement age in the previous year are far more likely to be allocated a new teacher, and hence are far less likely to experience a repeat match. Overall, and using both empirical strategies, the findings of this chapter indicate that repeating a match increases student performance.

Furthermore, to inform about the effectiveness of looping as an education policy, this chapter also studies whether the benefits of looping aggregate up to the student, class, and school level. These levels of aggregation do not allow the use of age discontinuities, therefore, the estimates use fixed effects methods. The results are consistent with those at student-teacher level: repeat matches increase the performance of students, classes, and schools.

The chapter also explores potential channels through which looping may improve student outcomes. Specifically, it posits that repeat matches improve school climate and contribute to creating a positive learning environment. Using data from survey of teachers, the chapter shows that classes with more student-teacher repeat matches have higher attendance, teachers report better classroom behaviour and have higher expectations of their students' academic potential. The findings on the role of school and classroom climate as an underlying mechanism build on an extensive educational literature studying the effects of a positive school environment on student outcomes (e.g. Thapa et al., 2013; Kraft et al., 2016; Klugman, 2017).

The findings from this chapter have important implications. First, the results inform about a widely used and understudied education policy. In the case of Chile, over 50% of the students progressing from 7th grade to 8th grade have the same teacher in both grades. Second, one of the great practical advantages of looping is that only needs a re-assignment of existing teaching resources. Then, the use of looping might bring significant improvements in learning without incurring additional costs on schools.

The third chapter analyses the role of family networks in the development of children's human capital. In particular, it analyses the influence of siblings on children's educational outcomes. The identification of sibling spillovers is challenging due to the well-known problems with the estimation of peer effects, namely the reflection bias and the existence of unobserved correlated effects (Manski, 1993; Blume et al., 2011).

This chapter combines detailed administrative data from Chile to examine how children affect the long-term educational outcomes of their siblings. To causally identify sibling spillovers, it exploits an exogenous variation in children's school starting age caused by Chile's school entry cutoff date. Effectively, it compares the performance of children whose sibling was born just before the enrolment cutoff with those whose sibling was born just after.

This study starts by reporting the effect of own school starting age on student achievement. The findings confirm a well-documented result in the educational literature: children who enter school at an older age score higher on standardised tests, have higher school grades, and are more likely to attend college (e.g. Bedard and Dhuey, 2006; McEwan and Shapiro, 2008; Lubotsky and Kaestner, 2016; Gallegos and Celhay, 2020). In addition, this chapter contributes to the existing literature by showing that older siblings are less affected by the school starting age.

The second part of the chapter investigates the existence of sibling spillovers on college entrance exams generated by birth date cutoff rules. Previous studies have almost exclusively focused on spillovers from older to younger siblings (e.g. Joensen and Nielsen, 2018; Landersø et al., 2020; Altmejd et al., 2021). By contrast, this chapter studies spillovers in both directions, i.e. from younger-to-older siblings and from older-to-younger siblings. The findings reveal strong positive spillovers running from younger-to-older siblings. This counter-intuitive result means that older siblings benefit from younger siblings starting school at an older age. Also, the gains accrue only to students coming from high-income families, who are close in age to the younger sibling and have higher school grades before their sibling make the transition into school. Surprisingly, this advantage does not spillover to their younger siblings.

The chapter also explores channels through which spillovers run from younger-to-older siblings. Using survey data containing information about parental investments, it shows that older siblings receive less attention from their parents when their sibling starts school at a younger age. This finding builds upon the existing studies on intra-household resource allocation and its interaction with early life shocks on human capital (e.g. Rosenzweig and Zhang, 2009; Del Bono et al., 2012; Yi et al., 2015).

The present study adds to the growing body of research that emphasises the role of family networks in children's educational trajectories. Specifically, it shows that student learning outcomes are affected by their sibling's educational history (e.g. school starting age). Taken together, the results of this research support the hypothesis of joint formation of human capital within the family, and stress the importance of take into account spillover effects on the family, when designing, evaluating, and implementing educational interventions.

Chapter 1

Gender differences in grading: teacher bias or student behaviour?

1.1 Introduction

School grades are one of many tools that teachers use to provide feedback to students about their learning. However, teachers' evaluations conflate student cognitive performance with non-cognitive factors, such as student effort, engagement, and class behaviour; and therefore, may potentially reflect teachers' biases (Brookhart et al., 2016). In addition, previous research has established the direct influence of teachers' grading behaviour on student motivation, self-confidence, and effort (e.g. Figlio and Lucas, 2004; Bonesrønning, 2008; Mechtenberg, 2009; Burgess and Greaves, 2013).

In recent years there has been growing interest in whether gender-biased perceptions affect teachers' grading behaviour. This question has taken on significance in the context of widely known and documented gender gaps in student performance. Do teachers grade differently according to gender? Can grading gaps be interpreted as evidence of teachers' biases, or do they only reflect differences in student behaviour? This chapter addresses these two questions. I proceed in two steps. First, using precise and comprehensive student-level administrative data, I examine whether teachers' grading behaviour systematically differs across gender. Based on prior studies (Blank, 1991; Goldin and Rouse, 2000; Lavy, 2008), I use a difference-in-differences strategy to capture gender differences in teachers' grading. In particular, I compare school grades marked by teachers and national exams marked externally and blindly. Second, I explore whether these differences can be attributed to "bias" on the part of teachers, or differential behaviour on the part of children in their performance on different tests.

This chapter connects to the literature on gender bias in teachers' grading.[1] Usually, the research methodology in this literature is based on comparing blind and non-blind tests scores. This strategy exploits the fact that in the blind test the student's identity is not revealed – and therefore it is assumed free of any teacher bias – while in the non-blind test, the student's identity is known. Then, provided that both tests measure the same abilities, the blind score can be conceived as a counterfactual measure to the non-blind score, and so the difference between the two test scores corresponds to a measure of the teacher's bias.[2] Contrary to the general belief that teachers may be biased against female students (Tiedemann, 2000; Ceci et al., 2014), most of the studies have found that the gender gap is against male students. Teachers' pro-female bias has been documented in several countries and educational contexts, including Czech Republic (Protivínský and Münich, 2018), France (Terrier, 2020), Israel (Lavy, 2008; Lavy and Sand,

[1] There is also a related and burgeoning literature that investigates the consequences of gender bias in teachers' grading on long-term educational outcomes (Lavy and Sand, 2018; Lavy and Megalokonomou, 2019; Terrier, 2020). Lavy and Sand (2018) study the short and long-term effects of being exposed to biased teachers at primary school in Israel. They find that being assigned a teacher who over-assess girls (boys) in primary school in a particular subject has a positive impact on girls' (boys') future academic achievements in that subject in both middle and high school test scores. Also, a bias in favour of girls (boys) in math has a positive effect on girls' (boys') enrolment in math advanced studies in high school. Similarly, Terrier (2020), using a dataset of 35 middle schools in France verifies that having been exposed to a teacher who is biased against boys in 6th grade has a negative effect on boys' progress relative to girls between the beginning of 7th and the end of 9th grade. Finally, Lavy and Megalokonomou (2019) using panel data on teachers and students from 21 high schools in Greece, show that there is a negative relationship between teacher's valued added – which is a proxy of teacher's quality – and teachers' pro-female or pro-male bias.

[2] The differences between blind and non-blind assessments have been extensively used by discrimination literature, see for example Goldin and Rouse (2000) and Blank (1991). This strategy has been exploited to explore differences in teachers' assessments across ethnic groups (Burgess and Greaves, 2013) and racial groups (Botelho et al., 2015).

2018), Italy (Casula and Liberto, 2017), Norway (Falch and Naper, 2013) and United States (Cornwell et al., 2013).[3]

What might explain this observed gender bias? The literature has identified a number of explanations for the bias against boys. First, it may be driven by teachers practising statistical discrimination (Phelps, 1972; Arrow, 1973). Under this theory, teachers might use observable characteristics to proxy for unobservable ability. In this respect, Lavy (2008) using data from public high schools' teachers in Israel, argues that if teachers are influenced by the expected performance of the group, teachers should give higher marks to the sex which performs better in that school. However, he finds that regardless of the relative performance of boys and girls in a school, the bias is always against boys.

Second, previous research emphasises the role of teachers, exploring the relationship between gender bias and teachers' characteristics.[4] Lavy (2008) shows that the relationship between teacher characteristics (e.g. gender, age, experience, and number of children) and the gender bias varies from subject to subject.[5] In another example, Falch and Naper (2013) find that a higher proportion of female teachers at the school implies lower grades for female students in Norwegian language, whereas there is no effect in math and English. In a recent study, Lavy and Megalokonomou (2019), using panel data on teachers and students from 21 high schools in Greece, show that the teachers' biased behaviour is highly persistent across classes.

Third, several studies have shown that women are less effective compared to men in competitive environments (e.g. Gneezy et al., 2003; Gneezy and

[3] An exception is Hinnerich et al. (2011). They detect no evidence of grading bias by gender in the case of Sweden. However, they only test the difference between blind and non-blind test scores for one subject (Swedish). See Protivínský and Münich (2018) for a fuller overview of the literature on gender grading gap.

[4] Previous research has established the role of student-teacher interactions on different educational outcomes. In particular, they can have an important influence on gender differences (Sadker and Sadker, 1994; Dee, 2005). For example, research has highlighted that girls may benefit from being assigned to female teachers (Bettinger and Long, 2005; Hoffmann and Oreopoulos, 2009).

[5] For example, in math, only male teachers exhibit a bias against boys, whereas no gender difference in grading is observed in the case of female teachers. In English, the effect is reversed.

Rustichini, 2004; Niederle and Vesterlund, 2007; Shurchkov, 2012; Azmat et al., 2016). Consequently, boys may overperform in high-stakes test compared to girls, since it is a more competitive environment. Then, if the blind test is a high-stakes test, this behaviour may explain the bias in favour of girls. However, the basic finding that the bias is against boys does not vary with different levels of competitiveness (Falch and Naper, 2013; Casula and Liberto, 2017; Terrier, 2020).

Finally, it is conceivable that differences in students' behaviour explain the difference in teachers' evaluations. In particular, teachers – consciously or unconsciously – might reward positive behaviour by giving higher marks. Cornwell et al. (2013) using data on teachers' perception about classroom behaviour at student level show that when student behaviour is taken into account no gender bias against boys is found. One major drawback of this approach is that the measure of non-cognitive skills relies on the teacher's perception about the student' behaviour, which may also be biased.[6]

Together, the reason for teachers' grading bias in favour of female students remains unexplained. It does not appear to be the result of discrimination. The evidence on teacher characteristics is far from conclusive. There is no evidence that it results from anonymised tests taking place in a more competitive environment. Finally, the evidence that it results from students' behaviour relies on teachers' – possibly biased – perceptions.

In this chapter, I provide new and more definitive evidence on the extent and causes of grading gaps between non-anonymised and anonymised central test scores. As previous studies, I exploit the fact that the national test is anonymously marked, and therefore is a blind test; whereas the school grades are marked by the teacher who knows the identity of the student, and therefore it can be considered as a non-blind test. I show that deviations between school grades and standardise tests might reflect not only

[6] In other words, if a teacher dislikes students – because of their gender – she will perceive them to have a "bad behaviour"' and will also give them low grades. If so, controlling for teacher's perception will be capturing the bias and the grading effect will disappear. In addition, as mentioned by Protivínský and Münich (2018), their evidence is based on younger children, and therefore does not explore the role of students' behaviour in teachers' grading at older ages.

teacher's biases, but also differences in how these two types of evaluations are linked to factors related to students' achievement, such as student effort, ability, and behaviour. I use an administrative student-level dataset on school grades and national test scores in Chile for the period 2011–2018. This information is available for two subjects: Spanish language and math. One particular feature of the data is that I have information about these two types of assessments at four different times – 4th grade (students aged 9-10), 6th grade (students aged 11-12), 8th grade (students aged 13-14), and 10th grade (students aged 15-16) – which I show has a consistent and strong effect on the gap between blind and non-blind test scores. Additionally, using teachers' administrative records I am able to match students and their subject-teachers, as well as their characteristics, including gender, teaching experience, and working conditions.

I start by documenting the existence of gender differences in grading behaviour. Consistent with the literature, I find that boys get lower grades than girls when they are assessed by their teachers compared to their scores in the national exams, in both Spanish and math. This suggests that there is a grading gap, and this runs against boys. The grading gap against boys remains unchanged by school type, rural/urban schools, school size and geographical location. In addition, I provide evidence that this grading gap holds for all the grades examined. Moreover, it presents a clear pattern, increasing from 4th grade to 8th grade, before falling to the transition to 10th grade.

In addition, I construct a classroom-subject dataset to quantify how much of the variation in the grading gap by gender can be attributed to schools, teachers, and classes, controlling for observed characteristics of the teachers. This dataset allows to compute teachers' grading behaviour in different classrooms, with different groups of students during the period of study. In contrast to Lavy and Megalokonomou (2019), I show that teachers' grading behaviour is far from being a fixed characteristic of the teacher. In turn, it seems to be governed by the characteristics of the class and, more specifically, by the characteristics of the students.

Finally, by exploiting rich survey information, I provide evidence on the mechanisms that help to understand the grading gap against boys. Building from previous results and the characteristics of the tests, it is possible to rule out three of the potential mechanisms identified by the literature, namely, *statistical discrimination hypothesis*, *teachers' characteristics* and *competitive environment*. Consequently, I focus on the role of student behaviour at school in explaining the pro-female grading gap. To accomplish this, I include proxy variables for school effort – using administrative data of school attendance and grade retention – and students' attitudes to learning – using students' surveys. This alleviates the concern that these variables could pick up teachers' biased perceptions. Interestingly, when these variables are included, the grading gap against boys vanishes. In particular, female students experience higher (more positive) returns to behavioural variables than male students. This could be the case either by effort inputs being more valuable for girls than boys or by teachers' biased behaviour in rewarding that effort.

Overall, this chapter contributes to the existing literature in several ways. First, it contributes with new evidence of a grading gap against boys, in a middle-income country as Chile. To the best of my knowledge, it is the first evidence that the grading gap in favour of girls also applies in a middle-income country with relatively low-performing schools. Second, whereas most studies analyse the gender gaps in grading using limited data samples, this chapter uses student-level administrative data.[7] Using administrative data not only improves the precision of the estimates, but also allows to link students to teachers and schools. Third, this is the first study to evaluate whether the grading gap varies systematically with the age of the children, showing that the pro-female grading gap is present throughout all the school years, with a clear pattern increasing with age up to 8th grade and decreasing in the transition to 10th grade. Fourth, this study provides a more comprehensive understanding of sources of variation of the gaps in grading. In contrast to some earlier studies, it is concluded that teachers' grading behaviour is not persistent across classes nor is it independent of the classroom environment. Fifth, the findings of this research

[7] There are only two exceptions, Falch and Naper (2013) and Casula and Liberto (2017).

provide insights into the explanations behind teachers' grading behaviour. In particular, add to the literature by emphasising the role of the students' behaviour in shaping gender differences in teachers' grading.

The remainder of the chapter proceeds as follow. Section 1.2 gives a brief description of the institutional setting and the data. Section 1.3 presents a theoretical motivation of the gender differences in grading. Section 1.4 explains the econometric framework and tests the existence of these differences. Section 1.5 quantifies and identifies the sources of variation of the gender differences in grading. Section 1.6 analyses the possible mechanisms behind the previous estimates. Finally, Section 1.7 offers a summary of the main findings and discusses their implications.

1.2 Institutional setting and data

1.2.1 The Chilean school system

The school system in Chile is organised in three levels: pre-primary education (up to 5 years old); primary education (6 to 13 years old); and secondary education (14 to 17 years old). Primary education consists of eight grades and is divided into two cycles: the first cycle – years 1 to 4 – and the second cycle – comprising years 5 to 8. Secondary education consists of four grades (years 9 to 12) and is structured in two cycles. The first two years offer general education, while the last two years involve a choice between academic studies and technical-professional/vocational studies. For primary and secondary education there are three types of school providers: (1) municipal or public schools, which are administered by the municipalities, and are financed through a per-student subsidy from central government; (2) private subsidized or voucher schools, which are administered by for-profit or non-profit private organizations, and receive a per-student subsidy as well as municipal schools. In addition, they can be financed through co-payment and unlike the public schools they can select their students; (3) private schools, which are run by private organizations (whether or not for profit) and receive no public funding.[8]

[8] For a detailed description of the Chilean school system, see Santiago et al. (2017).

1.2.2 School grading system and SIMCE test

The non-blind test score is based on school grades. The academic qualification in Chile uses a rating scale of 7 points with an increment of 0.1, 1 being the lowest and 7 the highest. Teachers set the final grade based on tests taken during the academic year between March and December. The evaluation standards and methods are decided autonomously by each school and/or teacher. School grades are relevant to be promoted to the next year.[9] In particular, high school grades are part of the eligibility criteria for admission to higher education.

On the other hand, the anonymised test score – which is not affected by teacher bias – is the *Sistema de Medición de la Calidad de la Educación* (SIMCE) test. This test corresponds to a standardised test administered by the Ministry of Education to all students in some particular grades, and it is the main instrument to measure the quality of education in Chile. The SIMCE test has no direct consequences for individual students' future prospects, therefore it is a low-stakes test. The test is administrated by external examiners and provides information about students' performance relative to the country's National Curriculum Framework. The test measures the areas of language (Spanish) and mathematics, and for some cohorts also measures the areas of natural sciences and social sciences.[10]

The SIMCE test is designed based on the measurement model of Item Response Theory (IRT), which is extensively used to assess student learning in most international large-scale standardised tests, such as PISA, TIMSS and PIRLS. Scores are scaled so that the national mean in each year is 250

[9] The minimum passing grade is 4.0 on average. There are other causes of repeating the year, the most prevalent is scoring below 4.0, on two or more subjects. Students who have two subjects below 4.0 can be promoted to the next year, as long as their average across all subjects is greater or equal to 5.0. Students in 1st and 2nd year in primary school are automatically promoted to the next year, as long as they meet a minimum attendance rate of 85%.

[10] Initially, students took the SIMCE test in 4th, 8th and 10th grade, although the test has expanded over the years so that students take it more frequently. Since 2012 tests also cover 2nd (only in Spanish) and 6th grade. However, in 2016 a reform was carried out in the schedule of the SIMCE test, which resulted in a reduction in the number of census-based assessments. As of 2016, the SIMCE test will be held each year for 4th and 10th grade students, while every second year, in alternate years, 6th and 8th grade students will be evaluated.

points, with a standard deviation of 50 points. The SIMCE test consists of both closed-ended question (multiple choice) and open-ended questions. Both types have a completely different marking process. Regarding the first type, the questions are captured and then digitised. The people who execute these tasks have access to the information contained in the answers sheet, such as the name of the school, the name of the student and the marks of the answers. However, they do not know what the correct answers are. The correction of open questions is carried out by means of a software that shows the corrector an image of the answer written by a student that is not individualised, thus it protects the personal and sensitive data of the students (including student's gender). For all practical purposes therefore, the SIMCE test is an anonymous and unbiased evaluation. National test results are public at national level (including regions and municipalities or districts) as well as at school level. The SIMCE test scores at individual or class levels are only available for researchers. Children, parents, teachers, and school directors do not have access to this information. Students take the SIMCE test before they know their final grades in school. Specifically, students take the SIMCE test in October/November, whereas students know their school grade usually at the end of the academic year.

1.2.3 Data

To study the gender differences in grading, I use data from students' achievements in the SIMCE test, as well as data from teachers' assessments, in the subjects of Spanish and math. The data contain unique school and *class* identifiers. In Chilean schools, a class is a group of students that take all the subjects together with exactly the same teachers, in a specific year and grade. This information allows me to aggregate the data to the class and school level.

The data on school grades are only available for a few years (2011–2018), whereas the SIMCE test is taken every year but has a specific grade schedule determined by the Ministry of Education. Therefore, the match between school grades and SIMCE test scores is feasible for 25 year-grade combinations (see Table 1.1). The school grading scale is from 1 to 7, while the

SIMCE test has a mean of 250 and a standard deviation of 50. Therefore, to make the tests comparable, school grades and SIMCE test scores are standardised to a distribution with zero mean and unit variance.

As I am interested in comparing male and female students within a class, the sample is restricted to mixed schools, so single-sex schools are dropped.[11]

Table 1.1. Data available for both school grades and SIMCE test scores

Year/Grade	4th grade	6th grade	8th grade	10th grade
2011	Yes	-	Yes	-
2012	Yes	-	-	Yes
2013	Yes	Yes	Yes	Yes
2014	Yes	Yes	Yes	Yes
2015	Yes	Yes	Yes	Yes
2016	Yes	Yes	-	Yes
2017	Yes	-	Yes	Yes
2018	Yes	Yes	-	Yes

I link the data on school grades and SIMCE test scores, to three datasets provided by the Ministry of Education. First, I use complete enrolment records of all students in Chile from 2008 onwards. These records include information on school characteristics (e.g. school type and total enrolment), individual school attendance, ethnicity, nationality, and whether the student passed or failed that school year. Second, I use SIMCE Complementary Survey, which are answered by students and parents. The first questionnaire registers students' beliefs and attitudes about studying and learning. From parents' surveys, it is possible to obtain information about students' socio-economic background, such as mother's and father's education and income (hundred thousand CLP). I complement this information with the SIMCE test scores in 4th grade. Finally, as a third source of information, I use data from teachers' administrative records to identify teachers of Spanish and math. This dataset also contains information on teacher gender, teaching experience and working conditions. More impor-

[11] Single-sex schools are not common in Chile. Between 2011 and 2018, about 98% of the schools in Chile were mixed schools.

tantly, this dataset allows to match students and their teachers in each subject, and to follow teachers over time, across classes and grades.

I restrict the analysis to students with valid school grades and SIMCE test scores in both subjects, and a complete set of information on characteristics. As a result, the sample comprised 2,821,911 students, 40,044 teachers, 111,074 classes and 5,126 schools. Table 1.2 presents summary statistics. Panel (a) reports information at the student level. For each student, I observe sex, family background, indicators of ethnicity and foreign student, SIMCE test scores in grade 4, *school attendance* and *grade retention*. School attendance corresponds to the average attendance over the previous three years, whereas grade retention is an indicator variable that takes value of 1 if the student failed in at least one of the previous three school years. Panel (b) shows information at the teacher level, including sex, average years of experience, and working conditions. Information at class level is shown in Panel (c), which includes class size, average number of boys and average number of girls. Panel (d) presents information at school level, such as average enrolment over 2011–2018, school type (i.e. public, voucher or private), urban schools, geographical location, and number of classes within the school.[12]

1.3 Measuring teacher bias in theory and practice

For each student i, it is observed a school grade G_i and a SIMCE test score T_i. The key difference between these two assessments is the nature of the scoring process. On the one hand, the school exams are marked by the teacher, who observes the identity of the students, including their gender. Then, school exams can be considered as a non-blind evaluation (NB). On the other hand, the SIMCE test is anonymously graded by external examiners, and consequently, it is a blind evaluation (B).

[12] Chile is administratively divided into 15 regions, including the Santiago Metropolitan Region where the national capital Santiago is located.

Table 1.2. Descriptive statistics

	All students		4th grade		6th grade	
	Mean	SD	Mean	SD	Mean	SD
(a) Student level						
1=Female	0.50	(0.50)	0.50	(0.50)	0.49	(0.50)
Father's education	12.11	(3.67)	12.40	(3.58)	12.20	(3.62)
Mother's education	12.21	(3.49)	12.54	(3.39)	12.32	(3.44)
Household income	6.14	(5.76)	6.65	(5.98)	6.33	(5.79)
1=Ethnic group	0.05	(0.22)	0.05	(0.21)	0.05	(0.22)
1=Foreign student	0.004	(0.06)	0.004	(0.06)	0.006	(0.08)
4th grade score	0.31	(1.01)			0.33	(0.99)
School attendance	93.31	(4.87)	92.82	(5.07)	93.38	(4.75)
1=Grade retention	0.07	(0.25)	0.05	(0.23)	0.06	(0.24)
Observations	2,821,911		931,129		577,761	
(b) Teacher level						
1=Female	0.76	(0.43)	0.87	(0.34)	0.77	(0.42)
Experience	13.37	(11.58)	13.71	(11.18)	13.25	(11.79)
1=Permanent contract	0.57	(0.45)	0.63	(0.45)	0.55	(0.47)
Working hours	37.44	(5.74)	37.45	(5.16)	37.46	(6.04)
Observations	40,044		16,906		16,595	
(c) Class level						
Class size	25.41	(6.68)	26.37	(6.96)	25.22	(6.68)
Number of boys	12.78	(4.37)	13.31	(4.26)	12.76	(4.11)
Number of girls	12.63	(4.73)	13.06	(4.57)	12.46	(4.50)
Observations	111,074		35,305		22,908	
(d) School level						
Enrolment	562.23	(400.82)	576.27	(410.00)	587.44	(413.34)
1=Public	0.45	(0.50)	0.41	(0.49)	0.41	(0.49)
1=Voucher	0.49	(0.50)	0.53	(0.50)	0.53	(0.50)
1=Private	0.06	(0.23)	0.07	(0.25)	0.06	(0.24)
1=Urban	0.89	(0.31)	0.90	(0.30)	0.90	(0.30)
1=Metropolitan	0.34	(0.47)	0.36	(0.48)	0.35	(0.48)
Number of classes	21.67	(18.67)	8.48	(6.41)	5.74	(4.11)
Observations	5,126		4,161		3,994	

TABLE 1.2. Descriptive statistics (continued)

	8th grade		10th grade	
	Mean	SD	Mean	SD
(a) Student level				
1=Female	0.50	(0.50)	0.50	(0.50)
Father's education	11.89	(3.73)	11.84	(3.74)
Mother's education	11.97	(3.54)	11.90	(3.56)
Household income	5.78	(5.56)	5.62	(5.52)
1=Ethnic group	0.05	(0.22)	0.05	(0.23)
1=Foreign student	0.004	(0.06)	0.003	(0.05)
4th grade score	0.27	(1.03)	0.32	(1.02)
School attendance	93.67	(4.64)	93.59	(4.84)
1=Grade retention	0.07	(0.25)	0.10	(0.30)
Observations	587,549		725,472	
(b) Teacher level				
1=Female	0.70	(0.46)	0.63	(0.48)
Experience	12.68	(11.66)	11.84	(10.95)
1=Permanent contract	0.53	(0.47)	0.51	(0.46)
Working hours	37.36	(6.43)	37.82	(6.81)
Observations	17,356		12,825	
(c) Class level				
Class size	24.84	(6.49)	24.84	(6.36)
Number of boys	12.52	(4.02)	12.36	(4.87)
Number of girls	12.32	(4.50)	12.48	(5.23)
Observations	23,656		29,205	
(d) School level				
Enrolment	594.34	(415.05)	749.39	(464.62)
1=Public	0.44	(0.50)	0.29	(0.46)
1=Voucher	0.50	(0.50)	0.60	(0.49)
1=Private	0.06	(0.24)	0.10	(0.30)
1=Urban	0.90	(0.30)	0.95	(0.23)
1=Metropolitan	0.34	(0.47)	0.35	(0.48)
Number of classes	5.77	(4.11)	12.07	(10.85)
Observations	4,101		2,420	

Notes: This table displays the means and standard deviations (SD) of the main variables used in this study. Sample comprises students in the year-grade combinations in which data of school grades and SIMCE test scores are available (see Table 1.1).

Hence, a true measure of teacher bias would be a comparison between school grades and SIMCE test scores under the two marking schemes, that is to say, $G_i(NB) - G_i(B)$ and $T_i(NB) - T_i(B)$. However, it is not possible to observe $G_i(B)$ or $T_i(NB)$. An alternative is to use $T_i(B)$ and consider – as a measure of bias – the difference between school grades and the SIMCE test scores: $G_i(NB) - T_i(B)$ (henceforth referred to as $G_i - T_i$). Therefore, the reliability of this measure of bias depends on the extent to which the SIMCE test score is a counterfactual for the score that the student would have received in the school if the school grade was blind. Suppose that school grades G_i are determined by the following function:

$$G_i = g(a_i, e_i^G) + b_{ij} + u_i^G \tag{1.1}$$

The specification captures the idea that the school grade is a function of the student ability a_i, the student effort e_i^G, a teacher-level component b_{ij} (which may vary by student), and a random component u_i^G. The teacher-level component is a measure of how easy it is for student i to obtain a grade by a given teacher j.

On the other hand, the standardised test score T_i, is also a function of student ability a_i and student effort in the test e_i^T, and a different (independent) random component u_i^T; but does not depend on b_{ij} (since it is not marked by the teacher):

$$T_i = t(a_i, e_i^T) + u_i^T \tag{1.2}$$

Thus, the average difference between the school grades and the SIMCE test scores would capture the teacher-level component b_{ij}, as well as the difference between the two functions $g(\cdot)$ and $t(\cdot)$:[13]

$$E[G_i - T_i] = E[g(a_i, e_i^G) - t(a_i, e_i^T)] + E[b_{ij}] \tag{1.3}$$

The measure of gender gap in grading would be given by the comparison

[13] It is important to note that Equation 1.3 does not account for other factors that might affect $g(\cdot)$ and $t(\cdot)$, such as teacher quality, parental motivation or teachers' expectations. These factors will not affect the gap between the tests as long as they affect test scores equally for $g(\cdot)$ and $t(\cdot)$.

of this difference for girls and boys, i.e. $E[G_i - T_i|\text{girl}] - E[G_i - T_i|\text{boy}]$:

$$\begin{aligned} &= \underbrace{E[g(a_i, e_i^G) - t(a_i, e_i^T)|\text{girl}] - E[g(a_i, e_i^G) - t(a_i, e_i^T)|\text{boy}]}_{\text{Differences in production functions}} \\ &+ \underbrace{E[b_{ij}|\text{girl}] - E[b_{ij}|\text{boy}]}_{\text{Gender bias in grading}} \end{aligned} \tag{1.4}$$

Equation 1.4 would be a reliable counterfactual measure of gender bias in grading only if $g(\cdot) = t(\cdot)$. More generally, it captures not only about differences in teacher behaviour captured by the term $E[b_{ij}|\text{girl}] - E[b_{ij}|\text{boy}]$, but also about differences in $g(\cdot) - t(\cdot)$ between girls and boys. For instance, school grades might measure different types of skills compared to the SIMCE test scores. This could be the case if school grades are based on homework, while the SIMCE test is a single test, and student performance in homework differs by gender. However, even if both functions are equal, there might be differences in the production inputs. For example, a student might try harder at the school vis-à-vis the SIMCE test, because it comes with praise from the teacher, or a student might try less hard because effort to achieve school grades is more costly than effort to achieve the SIMCE test score.

The teacher-level component b_{ij} might be a function of the students' behaviour $b_{ij}(X_i)$. For instance, teachers' grading behaviour might be a reaction to student engagement in the classroom. Then, student engagement in the classroom has two effects: firstly, it increases both the school grades and the SIMCE test score, even though the size of the effect might be different for each evaluation; and secondly, it might cause the teacher to increase their school grades, as a response to this behaviour. But, in order to observe a gender bias in grading, teachers should weight differently the student engagement depending on the gender of the student. Otherwise, there will be a bias, but it will be the same for girls and boys.

The framework described in this section helps understand the potential explanations identified in the literature behind the gender gap in grading. For instance, teachers' taste-based discrimination will be captured by the gender bias in grading: $E[b_{ij}|\text{girl}] - E[b_{ij}|\text{boy}]$. In this case, teachers have preferences for students of a particular gender, and express their preferences through grading. The statistical discrimination hypothesis could also be analysed by extending the model, allowing b_{ij} to be a function of the teachers' beliefs about the average performance of the student gender and the actual result on the school exam. As discussed earlier, teachers' grading behaviour might also be explained by their characteristics. In this case, the teacher-level component will be a function of the teacher characteristics $b_{ij}(Z_j)$. In addition, it is possible to include student-teacher interactions $b_{ij}(X_i, Z_j)$ to allow the presence of in-group bias, that is, teachers discriminate in favour of their own group. On the other hand, gender differences in attitudes towards competition might impact student effort allocation across the tests (e_i^G and e_i^T), which in turn, will affect their performance on each test. Similarly, gender differences in behavioural problems could be captured by the effort component, which affect both school grades and SIMCE test scores.

Altogether, this framework is able to account for all the potential explanations discussed in the literature. The next section presents the empirical methods and the estimation results, using Equation 1.4 as a definition of gender gap in grading.

1.4 The gender gap in teacher grading

As an initial approximation to the problem, Tables 1.3 and 1.4 show the results obtained from a mean-difference test of school grades (non-blind test) and SIMCE test scores (blind test) by gender, in Spanish and math, respectively. This comparison is quite revealing in several ways. First, there is a gender gap in both school grades and SIMCE test scores. In school grades, girls outscore boys in both subjects, but the difference is larger in Spanish than math. On the other hand, in the SIMCE test girls outperform boys in Spanish, while boys outperform girls in math. Furthermore, in absolute terms and for both tests, the gender gap is larger in Spanish and tends to increase as children get older. Second, Column 3 reports the difference between non-blind and blind scores. Therefore, girls tend to receive higher school grades compared to their SIMCE test score in both

Spanish and math, while the reverse is true for boys. Third, the last element of the Column 3 presents the difference-in-difference estimator, which corresponds to the difference between the non-blind and blind scores for girls minus the same difference for boys. Thus, a positive number indicates that the gender gap is in favour of girls; a negative number indicates that the gender gap is in favour of boys. The standard difference-in-differences estimation (DD) ranges between 0.08-0.18 standard deviations in Spanish and 0.08-0.22 standard deviations in math, respectively. The DD estimate increases between 4th grade and 8th grade, and then falls in 10th grade.

The DD estimate can also be expressed in terms of the following standard double-differences equation:

$$G_i - T_i = \beta + \gamma F_i + \zeta_i \tag{1.5}$$

where G_i and T_i are the school grade (non-blind test) and the SIMCE test score (blind test) for student i, respectively; and F_i is a female dummy. The constant β is informative of teachers toughness, also called grading inflation. The parameter γ measures the gender differences in scores gap - expressed in standard deviations - between male students and female students. Thus, a positive (negative) value for γ can be interpreted as evidence of a grading gap against boys (girls). From Equation 1.5, it is not possible to observe the part of the gender gap that does not depend on the type of test. To address this concern, Lavy (2008) proposes to run the following regression:

$$Y_{is} = \kappa + \alpha F_i + \beta NB_{is} + \gamma(NB_{is} \times F_i) + u_{is} \tag{1.6}$$

where Y_{is} is the score for student i under the scheme s, which can be blind (B) or non-blind (NB); F_i is a dummy for female student; and NB_{is} is equal to one if the test is non-blind and zero otherwise. So, an additional coefficient is displayed. The parameter α that represents the gender gap in the blind test and constitutes the gender gap that is common for both types of tests.[14]

[14] Because the balanced panel nature of the data – where every student has two observations $s = \{B, NB\}$ – this model is identical to estimating Equation 1.6 with student fixed effects. Consequently, the estimation of the parameter γ will not be

TABLE 1.3. Gender test score gap, by evaluation scheme: Spanish

		School grades	SIMCE test	Difference
		(non-blind)	(blind)	
		(1)	(2)	(3)
4th grade	Female	0.136	0.094	0.042
		(0.001)	(0.001)	(0.001)
	Male	-0.134	-0.092	-0.041
		(0.001)	(0.001)	(0.001)
	Differences	0.270	0.186	0.083
		(0.002)	(0.002)	(0.002)
6th grade	Female	0.187	0.106	0.081
		(0.002)	(0.002)	(0.002)
	Male	-0.183	-0.103	-0.079
		(0.002)	(0.002)	(0.002)
	Differences	0.370	0.209	0.161
		(0.003)	(0.003)	(0.002)
8th grade	Female	0.194	0.106	0.088
		(0.002)	(0.002)	(0.002)
	Male	-0.191	-0.104	-0.087
		(0.002)	(0.002)	(0.002)
	Differences	0.385	0.209	0.175
		(0.003)	(0.003)	(0.003)
10th grade	Female	0.197	0.138	0.059
		(0.002)	(0.002)	(0.002)
	Male	-0.199	-0.139	-0.060
		(0.002)	(0.002)	(0.002)
	Differences	0.396	0.278	0.118
		(0.002)	(0.002)	(0.002)

Notes: Standard errors are reported in parentheses.

The results of Equation 1.6 are shown in Table 1.5. The gender gap common to both tests (α) is positive in Spanish, and ranges between 0.19-0.28 standard deviations. On the contrary, in math, the gender gap is negative and ranges between 0.07-0.14 standard deviations. On the other hand, and for both subjects, the teacher grading effect β is negative. This suggest that teachers tend to give lower grades compared to the SIMCE test. Fi-

affected by the inclusion of students' characteristics and/or class fixed effects. It will also have no impact on the grading inflation coefficient β. Put differently, the regression model assumes homogeneous effect of students' characteristics on both, blind and non-blind test scores. However, through the inclusion of additional variables and their interaction with the non-blind dummy, the regression model allows for different effects of these variables on blind (SIMCE test) and non-blind (school grades) test scores. Finally, Equation 1.6 is mathematically equivalent to Equation 1.5: $G_i - T_i \equiv Y_{iNB} - Y_{iB} = \beta + \gamma F_i + (u_{iNB} - u_{iB}) = \beta + \gamma F_i + \zeta_i$.

Table 1.4. Gender test score gap, by evaluation scheme: Math

		School grades (non-blind) (1)	SIMCE test (blind) (2)	Difference (3)
4th grade	Female	0.008	-0.034	0.042
		(0.001)	(0.001)	(0.001)
	Male	-0.008	0.033	-0.041
		(0.001)	(0.001)	(0.001)
	Differences	0.016	-0.068	0.084
		(0.002)	(0.002)	(0.002)
6th grade	Female	0.068	-0.033	0.101
		(0.002)	(0.002)	(0.002)
	Male	-0.067	0.032	-0.099
		(0.002)	(0.002)	(0.002)
	Differences	0.135	-0.065	0.200
		(0.003)	(0.003)	(0.002)
8th grade	Female	0.041	-0.071	0.112
		(0.002)	(0.002)	(0.002)
	Male	-0.040	0.070	-0.110
		(0.002)	(0.002)	(0.002)
	Differences	0.081	-0.141	0.222
		(0.003)	(0.003)	(0.003)
10th grade	Female	0.032	-0.045	0.077
		(0.002)	(0.002)	(0.002)
	Male	-0.032	0.046	-0.078
		(0.002)	(0.002)	(0.002)
	Differences	0.064	-0.091	0.155
		(0.002)	(0.002)	(0.002)

Notes: Standard errors are reported in parentheses.

nally, and in line with the results in Tables 1.3 and 1.4, γ is positive in Spanish and math, meaning that the grading gap is against boys.[15]

An important concern about the grading differences shown above, is that they may be affected by other characteristics of the students. To address this issue, Lavy (2008) proposes estimate the Equation 1.6 augmented by the interaction between the non-blind test and these characteristics:

$$Y_{is} = \kappa + \alpha F_i + \beta NB_{is} + \gamma (NB_{is} \times F_i) + \Phi (NB_{is} \times X_i) + \delta X_i + u_{is} \quad (1.7)$$

[15] In Appendix A.2 I estimate Equation 1.6 separately for low and high ability students, using 4th grade SIMCE test scores as a measure of ability. I show that the gender grading gap is positive (runs against boys) for low and high ability students.

Table 1.5. Estimation of gender gap in grading

	4th grade (1)	6th grade (2)	8th grade (3)	10th grade (4)
(a) Subject: Spanish				
Female	0.186***	0.209***	0.209***	0.278***
	(0.003)	(0.003)	(0.003)	(0.006)
Non-blind	-0.041***	-0.079***	-0.087***	-0.060***
	(0.008)	(0.009)	(0.010)	(0.014)
Non-blind × Female	0.083***	0.161***	0.175***	0.118***
	(0.002)	(0.003)	(0.004)	(0.007)
R-squared	0.013	0.023	0.024	0.029
Observations	1,862,258	1,155,522	1,175,098	1,450,944
(b) Subject: Math				
Female	-0.068***	-0.065***	-0.141***	-0.091***
	(0.003)	(0.003)	(0.004)	(0.008)
Non-blind	-0.041***	-0.099***	-0.110***	-0.078***
	(0.009)	(0.010)	(0.012)	(0.015)
Non-blind × Female	0.084***	0.200***	0.222***	0.155***
	(0.002)	(0.003)	(0.004)	(0.008)
R-squared	0.001	0.003	0.003	0.002
Observations	1,862,258	1,155,522	1,175,098	1,450,944

Notes: Standard errors are clustered at school level and are reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

where X_i is a vector of student characteristics, which includes socio-economic background characteristics (mother's education, father's education, and household income), demographic characteristics (indicators of ethnicity and foreign student) and the grade 4 SIMCE test score in the subject. The results of Equation 1.7 are presented in Tables 1.6 (Spanish) and 1.7 (Math). Firstly, the results show that the inclusion of the interaction between the non-blind test and student characteristics does not substantially change the gender gap in grading, and confirm those previously presented: the gender gap in grading runs against boys. Secondly, students from advantaged backgrounds (higher income and more educated parents) and with higher SIMCE test scores in 4th grade tend to receive lower school grades compared to SIMCE test score, while the opposite is observed for ethnic minority students and foreign students. Furthermore, these effects – with the exception of the SIMCE score in 4th grade – are substantially smaller than the gender interaction term.

Table 1.6. Estimation of gender gap in grading: Spanish

	4th grade (1)	6th grade (2)	8th grade (3)	10th grade (4)
Female	0.188***	0.076***	0.101***	0.189***
	(0.002)	(0.002)	(0.002)	(0.004)
Non-blind	-0.020	-0.008	0.195***	0.171***
	(0.015)	(0.016)	(0.016)	(0.025)
Non-blind × Female	0.083***	0.196***	0.209***	0.147***
	(0.002)	(0.003)	(0.004)	(0.007)
Non-blind ×				
Father's education	-0.004***	-0.004***	-0.010***	-0.008***
	(0.001)	(0.001)	(0.001)	(0.001)
Mother's education	0.003***	0.002**	-0.008***	-0.008***
	(0.001)	(0.001)	(0.001)	(0.001)
Household income	-0.001*	0.000	-0.004***	0.002
	(0.001)	(0.001)	(0.001)	(0.001)
Ethnic group	0.019	0.030**	0.046***	-0.021
	(0.013)	(0.014)	(0.014)	(0.020)
Foreign student	0.001	0.008	0.096***	0.089**
	(0.026)	(0.042)	(0.034)	(0.035)
Lagged SIMCE score		-0.175***	-0.186***	-0.166***
		(0.003)	(0.003)	(0.004)
R-squared	0.097	0.402	0.309	0.272
Observations	1,862,258	1,155,522	1,175,098	1,450,944

Notes: All regressions include student characteristics (father's education, mother's education, household income, indicators of ethnicity and foreign student) and lagged SIMCE test scores (except for 4th grade students). Standard errors are clustered at school level and are reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Figure 1.1 shows the value of γ for each of year-grade combination in which the data of school grades and SIMCE test scores are available (see Table 1.1).[16] As can be seen, it indicates that for all the combinations the gender gap in grading is against boys, and of the same order of magnitude than the pooled cross-sectional model. Interestingly, there is a clear pattern observable in the gender gap in grading: increasing through primary school (4th grade to 8th grade) and then decreasing in high school. Further checks are performed (see Appendix A.3). Particularly, the grading gap against

[16] Appendix A.1 reports the value of γ for each year-grade combination.

TABLE 1.7. Estimation of gender gap in grading: Math

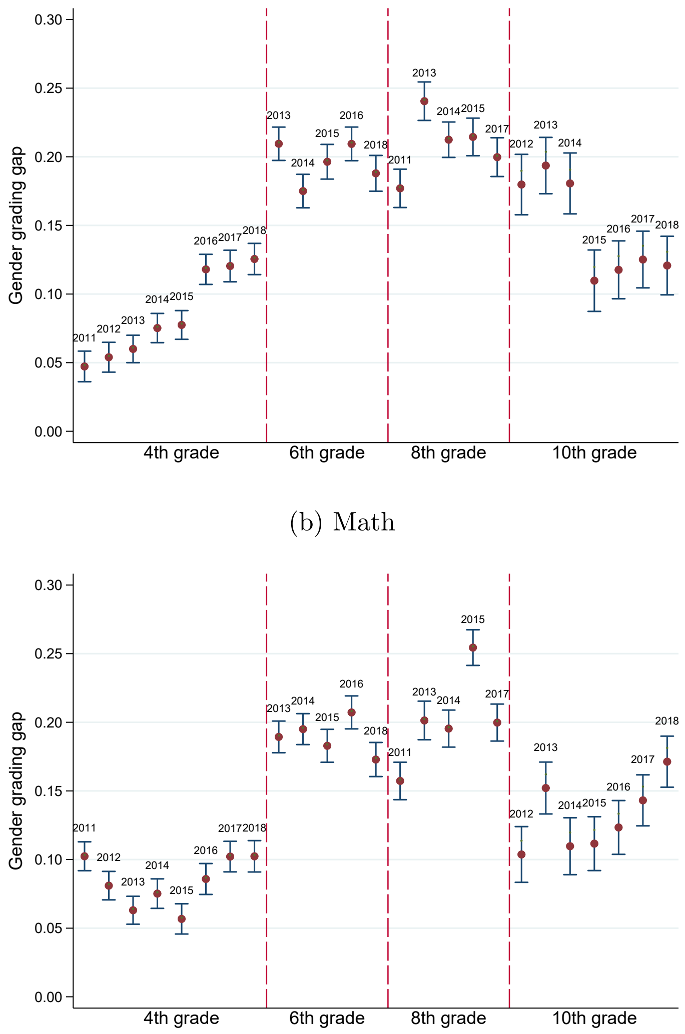
	4th grade (1)	6th grade (2)	8th grade (3)	10th grade (4)
Female	-0.066***	-0.022***	-0.083***	-0.028***
	(0.002)	(0.002)	(0.002)	(0.005)
Non-blind	0.203***	0.339***	0.531***	0.577***
	(0.017)	(0.017)	(0.018)	(0.025)
Non-blind × Female	0.083***	0.190***	0.202***	0.131***
	(0.002)	(0.003)	(0.004)	(0.007)
Non-blind ×				
Father's education	-0.008***	-0.012***	-0.019***	-0.020***
	(0.001)	(0.001)	(0.001)	(0.001)
Mother's education	-0.008***	-0.014***	-0.022***	-0.024***
	(0.001)	(0.001)	(0.001)	(0.001)
Household income	-0.008***	-0.012***	-0.019***	-0.016***
	(0.001)	(0.001)	(0.001)	(0.001)
Ethnic group	0.049***	0.101***	0.112***	0.050**
	(0.015)	(0.016)	(0.015)	(0.024)
Foreign student	0.083**	0.123***	0.187***	0.150***
	(0.034)	(0.047)	(0.044)	(0.033)
Lagged SIMCE score		-0.175***	-0.186***	-0.166***
		(0.004)	(0.004)	(0.005)
R-squared	0.081	0.437	0.355	0.302
Observations	1,862,258	1,155,522	1,175,098	1,450,944

Notes: All regressions include student characteristics (father's education, mother's education, household income, indicators of ethnicity and foreign student) and lagged SIMCE test scores (except for 4th grade students). Standard errors are clustered at school level and are reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

boys is robust across school type (public, voucher and private), urban and rural schools, school size and geographical region.

Overall, these results indicate that there is a gender gap in grading that runs against boys in both subjects. This finding is consistent with previous studies which have suggested that girls are graded more favourably than boys in all subjects. Despite this, the results differ from earlier studies, in at least three aspects. First, in math, the size of the effects differs from some of the earlier studies. The effects are more in line with the results of Casula and Liberto (2017), whereas are considerably larger than Lavy (2008) and Falch and Naper (2013). In turn, the grading gap against boys

Figure 1.1. Gender gap in grading using different year-grade combinations



(a) Spanish

(b) Math

Notes: The graph shows the estimates of γ in Equation 1.7 with 95% confidence interval for each year-grade combination.

in Spanish is considerably larger than the grading gap in native language documented by previous research. Second, and in contrast to previous studies, the magnitude of the grading gap against boys is similar for Spanish and math. Third, compared to most previous studies, the gender gap in grading is estimated more accurately.[17] Among other results, it is shown that the gender gap in grading is consistently positive (against boys) across

[17] An exception is Casula and Liberto (2017) estimations, which yield similar standard errors for their coefficients.

all year-grade combinations. In addition, the strong *grade effect* is suggestive that the gender gap is not only the result of teacher bias, but is also the result of student behaviour.

Are these results indicative of teachers' gender bias in grading? As shown in the previous section, these results might also be capturing differences in the production functions (or inputs to the production functions) of the school grades vis-à-vis the SIMCE test scores for girls and boys. In other words, they might capture any unobserved factor both correlated with gender and with test type. For example, observed differences between blind and non-blind test scores could be explained by the gender effort gap in the classroom, which differs from the gender effort gap in the SIMCE test. The following section provides evidence of the systematic impact of different factors on the gender gap in grading and clarifies to what extent this gap relies on teachers' behaviour and/or students' behaviour.

1.5 Does the gender grading gap capture teacher biases?

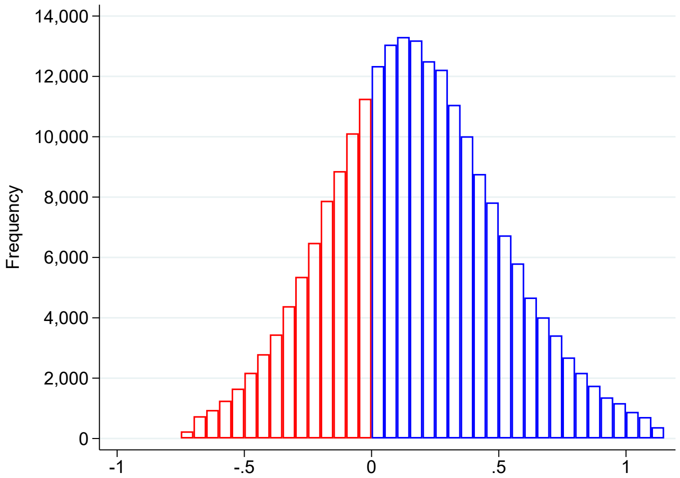
The previous section has presented robust evidence of a grading gap against boys in teachers' assessments. Throughout this section, I explore the extent to which the behaviour of relevant actors, such as schools and teachers, may explain this result. To identify the sources of variation in the gender gap in grading, I construct a dataset with information about the gender gap in grading at class-subject level, denoted by GG_{cs}. The gender gap in grading at class-subject level is defined as the difference between girls' and boys' gap between the average non-blind score (school grades) and blind score (SIMCE test score):

$$GG_{cs} \equiv (\overline{G}_{cs}^{F} - \overline{T}_{cs}^{F}) - (\overline{G}_{cs}^{M} - \overline{T}_{cs}^{M}) = \gamma_{cs} \tag{1.8}$$

This value is computed for each class-subject combination. Each class $c = \{1, ..., N\}$ is observed once in each subject $s = \{\text{Spanish}, \text{math}\}$. A class-subject combination has a specific teacher j, school k and year t.[18] As a result, it is obtained a dataset of 222,148 observations, composed of 40,044 teachers, 5,126 schools and 111,074 classes. On average, a teacher is observed in 5.5 different classes in the dataset (with a standard deviation of 3.4 times), whereas is observed 1.6 different classes within the same school year (with a standard deviation of 0.8 times). The average number of classes per school is 21.7 (with a standard deviation of 18.7). The gender gap at class-subject level has the same interpretation as before: a positive (negative) value for GG_{cs} implies a grading gap in favour of girls (boys).

Figure 1.2 plots the distribution of the gender grading gap at classroom-subject level. In coherence with previous results, on average the gender gap in grading is 0.17 (with a standard deviation of 0.37), and therefore runs against boys. It is worth noting that there exist classes where the gender gap in grading is negative, and therefore runs against girls.

Figure 1.2. Gender grading gap distribution at classroom-subject level



Notes: The figure omits observations below the first and above the 99th percentile of the gender gap in grading. Bins have width of 0.05 standard deviations.

Table 1.8 provides descriptive statistics. The gender grading gap at class-subject level mimics the results of the previous section, showing a clear pattern: increasing up to 8th grade and decreasing in the transition to

[18] GG_{cs} is equivalent to estimate Equation 1.6 for each class-subject (γ_{cs}).

10th grade. On the other hand, gender grading gap is higher in public schools, rural schools, and schools outside the Santiago Metropolitan Region.

Table 1.8. Descriptive statistics of gender grading gap at classroom-subject level

	Average	SD
(a) Subject		
Both subjects	0.17	(0.37)
Spanish	0.15	(0.39)
Math	0.19	(0.35)
(b) Grade		
4th grade	0.10	(0.31)
6th grade	0.21	(0.36)
8th grade	0.24	(0.40)
10th grade	0.17	(0.42)
(c) School characteristics		
1=Public	0.19	(0.41)
1=Voucher	0.16	(0.35)
1=Private	0.13	(0.35)
1=Urban	0.17	(0.37)
1=Rural	0.21	(0.40)
1=Metropolitan	0.16	(0.36)
1=Non-metropolitan	0.18	(0.38)

As grading is something that teachers do, it is particularly important to know the extent to which the gender grading gap is a fixed characteristic of a teacher. A particular feature of the data is that I can observe the same teacher in different classes. Are teachers' assessments always in favour of girls or boys? To characterise how teacher's grading change in different classes, I compute the number of class-subject combinations in which a class taught by a particular teacher, exhibits a grading gap against boys (i.e. the GG_{cs} is positive). Then, if the grading gap is a characteristic of the teacher, it should be observed that teachers are constantly biased in favour of girls or boys. This is not the case. About 23.4% of the teachers always present a grading gap against boys, whereas 2.7% of them always present a grading gap against girls. Accordingly, about one quarter of the teachers would exhibit a consistent bias.

Is teacher's grading gap stable? Following Lavy and Megalokonomou (2019), I measure the persistence of teachers' grading behaviour by comparing the gender gap of class c taught by teacher j ($\overline{GG}_{j,c}$), and the average gender gap over all the other classes taught by teacher j during 2011–2018, denoted by $\overline{GG}_{j,-c}$. This average captures all the information available about teacher j's grading behaviour, with different groups of students and classes. Moreover, the out–of–sample approach eliminates any class-level unobserved variation in boys' and girls' behaviour, along with any gender gap in non-cognitive skills. Table 1.9 presents the correlation between the gender gap of class c taught by teacher j and its out–of–sample average, using a school fixed effects regression with subject, year and grade fixed effects and controlling for teacher characteristics. In Column (1), I use the out–of–sample average; whereas in Column (2) the independent variable is the average weighted by the number of students ($\overline{GG}^{w}_{j,-c}$), to account for differences in class size. As a result, both measures produce a coefficient not statistically different from zero, meaning that teachers' grading behaviour is not persistent across classes.

Table 1.9. Regression of teacher's grading gap across classes

	(1)	(2)
Average other classes ($\overline{GG}_{j,-c}$)	0.003 (0.007)	
Weighted average other classes ($\overline{GG}^{w}_{j,-c}$)		0.009 (0.007)
R-squared	0.082	0.082
Observations	222,148	222,148

Notes: Dependent variable is the gender gap of class c taught by teacher j ($\overline{GG}_{j,c}$). All regressions control for teacher characteristics (gender, experience, an indicator of permanent contract and working hours) and include school fixed effects, subject fixed effects, grade fixed effects and year fixed effects. Standard errors are clustered at school level and are reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

To further test whether the gender gap in grading corresponds to a characteristic of the teacher, I follow the methodology proposed by Lavy and Sand (2018). They argue that if the measure of gender gap in grading is really capturing teachers' biased behaviour, it has to be the case that the correlation of the grading gap between subjects within the same class must be higher if both subjects are taught by the same teacher. In other words, under the hypothesis of gender gap in grading being an expression of the teacher's gender stereotypes, it is expected that a teacher persistently biases the same group of students, in both subjects. Table 1.10 shows the results of running a regression of grading gap in math on the grading gap in Spanish, for the same class. Column (1) presents the coefficient estimate between the two subjects (first row), and its interaction with an indicator variable for having the same teacher in both subjects (second row). Firstly, the results indicate that the grading gap is highly correlated across subjects (0.34 with a standard error of 0.003), even when the teachers are different. Secondly, the interaction coefficient is not statistically significant. Column (2) adds school fixed effects to account for school characteristics. No substantial changes are observed. The results suggest that the class is the most important element behind the gender gap in grading, and not the identity of the teachers.

TABLE 1.10. Regression of gender gap between subjects within the same class

	(1)	(2)
Gender grading gap in Spanish	0.337***	0.333***
	(0.003)	(0.003)
Gender grading gap in Spanish × Same teacher	0.010	0.009
	(0.008)	(0.008)
School FE		Yes
R-squared	0.178	0.249
Observations	111,074	111,074

Notes: Dependent variable is the gender grading gap in math. All regressions control for teacher characteristics (gender, experience, an indicator of permanent contract and working hours), and include grade fixed effects and year fixed effects. Standard errors are clustered at school level and are reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

What explains the variation in the gender gap in grading? The grading gap might show a different pattern depending on the subject, in particular, the set of skills assessed could differ between Spanish and math. Similarly, the age of the students could also be relevant, since gender differences in motivation at school and attitudes towards learning grow wider, as boys and girls grow and develop (Salisbury et al., 1999; Duckworth and Seligman, 2006; Kenney-Benson et al., 2006). Alternatively, as gender equality policies evolve, it is plausible to argue that the gender gap in grading may follow a particular trend over time. From an institutional point of view, it may respond to a school policy, for example, school directors may promote gender equality in the classrooms. Finally, the gender gap in grading could reflect a conscious or unconscious gender stereotype of the teacher. Then, to characterise how systematically the gender gap in grading changes due to these factors, I estimate the following regression model:

$$GG_{cs} = \mu_s + \mu_g + \mu_t + \mu_k + \mu_j + \beta Z_{jt} + \zeta_{cs} \tag{1.9}$$

where GG_{cs} is the gender gap in grading for the class c in the subject s; μ_s is a dummy for math; μ_g are grade fixed effects; μ_t are year fixed effects; μ_k are school fixed effects; μ_j are teacher fixed effects; and Z_{jt} is a vector of teacher's characteristics, including gender, years of teaching experience, an indicator of permanent contract, and working hours. It should be noted that teacher gender cannot be included with teacher fixed effects (μ_j). The idea is to use this model to test the explanatory power of different specifications, adding in each step a different set of variables.[19] Table 1.11 reports the results of estimate Equation 1.9. Regarding the coefficients, and across all the specifications, math teachers present a larger grading gap against boys compare to Spanish teachers. In addition, the gender gap in grading shows a clear pattern: it increases with age until the 8th grade, and then decreases until the 10th grade. On the other hand, no time trend is observed for the gender gap in grading over the years. Finally, there is very little systematic impact of the teachers' characteristics on the gender gap in grading.

Table 1.11 shows the F-test of overall significance of the fixed effects and its p-value, and the adjusted R^2. The first column exhibits the benchmark model which includes subject fixed effect (μ_j), grade fixed effect (μ_g), year fixed effects (μ_t), and the vector of teacher characteristics (Z_{jt}). The adjusted R^2 for this model is 2%. The second column adds school fixed effects

[19] A similar approach is used in (Bertrand and Schoar, 2003) to assess the impact of managers on business decisions.

to the benchmark model. Although the fixed effects are statistically significant, the adjusted R^2 reaches only 6%. Therefore, by itself, schools do not explain a substantial part of the variation of the gender gap in grading. In the third column, teacher fixed effects are added to the benchmark specification. Again, the fixed effects are jointly significant, but the model fit is 10%. As a result, teacher biased behaviour is not persistent over time, in other words, not much of the variation is explained by the teacher identity. No further changes are observed when school and teacher fixed effects are simultaneously included (fourth column). However, when class fixed effects are included, the adjusted R^2 jumps to 40%, besides being statistically significant.[20] In other words, a group of children – a class – who have a particular gender gap in Spanish have a similar gender gap in math. This confirms the results in Table 1.10. Finally, no major differences are observed when teacher fixed effects and class fixed effects are jointly included (adjusted R^2 increases by 5 percentage points).

Altogether, the results presented in this section suggest that teachers' grading behaviour is not fixed, i.e. the gender gap in grading does not substantively correlate with teachers' identity. In turn, it seems to be that the class characteristics – expressed in class fixed effects – are the key to understand the mechanism behind the gender gap in grading. In terms of the theoretical motivation discussed in Section 1.3, the results support the idea that the comparison between school grades and SIMCE test scores is capturing differences in production functions of girls and boys, rather than actual teachers' gender biases.

Nevertheless, in a more complicated setting, these results still might be driven by teachers' biases which depend on the students' behaviour. In other words, these outcomes do not necessarily imply that teachers play no role in this respect. For example, the grading gap against boys could be explained by teachers rewarding girls' *effort* more than boys' *effort*. The next section, therefore, moves on to discuss the possible mechanisms that support these findings, with a particular focus on the role of students' behaviour in shaping the grading gap by gender.

[20] It should be noted that class fixed effects saturate the effects at grade (μ_g) and year (μ_t) level.

Table 1.11. Stepwise regression: variation of gender grading gap

	(1)	(2)	(3)	(4)	(5)	(6)
Subject FE						
Math	0.035***	0.035***	0.018***	0.017***	0.035***	0.011***
	(0.002)	(0.002)	(0.003)	(0.003)	(0.001)	(0.003)
Grade FE						
6th grade	0.101***	0.098***	0.091***	0.088***		
	(0.002)	(0.002)	(0.004)	(0.004)		
8th grade	0.140***	0.138***	0.133***	0.129***		
	(0.002)	(0.002)	(0.004)	(0.004)		
10th grade	0.072***	0.086***	0.093***	0.088***		
	(0.002)	(0.003)	(0.006)	(0.006)		
Year FE						
Year 2012	0.026***	0.024***	0.023***	0.022***		
	(0.004)	(0.004)	(0.005)	(0.005)		
Year 2013	0.037***	0.036***	0.031***	0.032***		
	(0.003)	(0.003)	(0.004)	(0.004)		
Year 2014	0.027***	0.028***	0.020***	0.022***		
	(0.003)	(0.003)	(0.004)	(0.004)		
Year 2015	0.029***	0.032***	0.023***	0.025***		
	(0.003)	(0.003)	(0.004)	(0.005)		
Year 2016	0.037***	0.038***	0.026***	0.028***		
	(0.004)	(0.004)	(0.005)	(0.005)		
Year 2017	0.030***	0.033***	0.020***	0.023***		
	(0.004)	(0.004)	(0.005)	(0.006)		
2018	0.034***	0.036***	0.020***	0.024***		
	(0.004)	(0.004)	(0.006)	(0.006)		
Teacher characteristics						
Female	0.00186	0.00185			0.00073	
	(0.002)	(0.002)			(0.002)	
Experience	-0.00004	-0.00047***	0.00093*	0.00061	-0.00075***	0.00034
	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)	(0.001)
Permanent contract	-0.01745***	-0.00725***	0.00419	0.00390	-0.00859***	0.00342
	(0.002)	(0.002)	(0.003)	(0.004)	(0.003)	(0.006)
Work hours	-0.00059***	-0.00038***	-0.00030	-0.00026	-0.00034**	0.00017
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
School FE		Yes		Yes		
Teacher FE			Yes	Yes		Yes
Class FE					Yes	Yes
F-test FE		2.68	1.43	1.44	2.28	2.20
p-value		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Adj. R-squared	0.024	0.060	0.095	0.103	0.393	0.452
Observations	222,148	222,148	222,148	222,148	222,148	222,148

Notes: Standard errors are reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

1.6 Potential mechanisms behind teachers' grading behaviour

The results from the previous section discard two potential mechanisms: *statistical discrimination* and *teacher characteristics*; as both mechanisms are naturally linked to teachers' identity. In addition, given the specific features of the SIMCE test, it is possible to rule out that the gender grading gap is explained by a more competitive environment. As mentioned before, the SIMCE test has no direct consequences for individual students' future prospects, therefore it is a low-stakes test. Moreover, both school exams and SIMCE tests take place in the same environment. This arguably rules out the possibility that the SIMCE test might cause higher levels of anxiety in students compared to the school tests. Finally, the time lag between blind and non-blind tests is relatively small. Students take the SIMCE test before they know their final grades in school. Usually, the SIMCE test is taken during the months of October and November, whereas students know their final grade usually at the beginning of December. This avoids time trends in learning, which in principle could be different for male and female students. The following describes how student's behaviour may explain the grading gap against boys in teachers' assessments.

As some authors point out, the teacher grading gap could be a reaction to students' attitudes towards learning (Cornwell et al., 2013). More importantly, several lines of evidence suggest that gender correlates with the level of non-cognitive abilities (Bertrand and Pan, 2013). In particular, it is well established that boys tend to have more behavioural and attention problems (Ready et al., 2005), less self-regulation (Matthews et al., 2009) and less self-discipline (Duckworth and Seligman, 2006; Kenney-Benson et al., 2006) than girls. From a theoretical point of view, gender differences in behaviour might explain the gender gap either because they raise the student's achievement on school grades $g(\cdot)$ more than on SIMCE test $t(\cdot)$, or because they cause the teacher to be more positively biased $b_{ij}(\cdot)$.

The survey data contain information on students' perception about their own school effort. 31% of the students in the sample have this information.[21] To characterise the student's behaviour, four variables are considered. The first two variables are based on students' perception. Students were asked to indicate the extent of their agreement with the following statements: "I always do my homework" and "I like to study". The first is considered a measure of student effort, whereas the second a measure of positive attitude towards learning. The rating scale is "I fully agree", "I agree"; "Disagree", "I entirely disagree". Both variables are coded as dummy variables, taking value of one if the student answers "I fully agree" or "I agree", and zero otherwise. Furthermore, two additional measures of school effort are added: grade retention and school attendance. Grade retention reflects the past academic performance and can be considered as a mix of direct measure of ability and school effort. On the other hand, the school attendance rate is a direct measure of school effort in the previous years.

Table 1.12 displays descriptive statistics for each of these variables by gender. From this data, it can be seen that female students present higher levels of effort and positive attitude, and less grade retention. In contrast, no substantial differences in school attendance are observed. It is worth noting that the behaviour gender gap, in terms of effort and positive attitude, tends to get wider as children get older.

To test the impact of student behaviour on the grading gap, I include these four behavioural variables (*Do homework*, *Like to study*, *Grade retention*, and *School attendance*) along with their interactions with gender and non-blind test in Equation 1.7. Table 1.13 shows the results. Column (1) displays the baseline regression based on Equation 1.7. Once again, the grading gap is against boys in Spanish and math.

Column (2) reports the effects of these variables and their interaction with the type of test. This model allows the coefficients to vary with the type of test. In the light of the theoretical model presented in Section 1.3, this

[21] Table A.9 (Appendix A.4) reports a mean comparison test of student, teacher and school observable characteristics for the estimation sample and the restricted sample.

Table 1.12. Gender gap in behaviour

		All students	6th grade	8th grade	10th grade
1=Do homework	Female	0.78	0.72	0.81	0.82
	Male	0.70	0.65	0.75	0.73
	Difference	0.07***	0.07***	0.06***	0.09***
		(0.001)	(0.002)	(0.002)	(0.002)
1=Like to study	Female	0.51	0.47	0.52	0.54
	Male	0.43	0.43	0.46	0.41
	Difference	0.08***	0.05***	0.07***	0.13***
		(0.001)	(0.002)	(0.002)	(0.002)
1=Grade retention	Female	0.06	0.05	0.05	0.08
	Male	0.09	0.08	0.08	0.11
	Difference	-0.03***	-0.04***	-0.03***	-0.03***
		(0.001)	(0.001)	(0.001)	(0.001)
School attendance	Female	93.66	93.57	93.69	93.73
	Male	93.51	93.40	93.38	93.77
	Difference	0.14***	0.18***	0.31***	-0.04**
		(0.010)	(0.016)	(0.020)	(0.017)
Observations		870,765	347,145	235,470	288,150

Notes: Standard errors are reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

could be caused by two alternative mechanisms, which in principle are not mutually exclusive. On the one hand, it would capture differences in the production function, that is to say, same effort inputs $e = e^*$ would be more valuable in one test than the other: $g(e^*) \neq t(e^*)$. On the other hand, it could be capturing teachers rewarding effort inputs: $b_j(e_a) > b_j(e_b)$ with $e_a > e_b$.

The results show different effects according to the type of test. In particular, the effects of the behavioural variables are larger for the school tests than the SIMCE test. Also, they show that part of the grading gap against boys is explained by the differences in the behaviour by gender. The inclusion of these variables reduces the gender gap in grading by around 20%. However, the baseline grading gap is still positive (or against boys) and statistically significant. These results differ from Cornwell et al. (2013), who show that gender differences between external test scores and teachers' assessments vanish when behavioural variables are considered.

Column (3) presents estimates adding a triple interaction between behavioural variables, the non-blind test dummy, and the female dummy. A positive coefficient implies a higher premium in the non-blind test (school grades) for female students who exert effort. The most remarkable result to emerge from this model is that grading gap against boys completely disappears for both subjects. The results suggest that the premium due to good behaviour is gender-dependent, and runs in favour of girls.

In terms of the theoretical model, this implies that $E[g(\cdot) - t(\cdot)|\text{girl}, e] > E[g(\cdot) - t(\cdot)|\text{boy}, e]$. This might be explained by two reasons, which are not mutually exclusive. First, it could reflect effort inputs being more valuable for girls than boys. It could also be argued that this effect is due to effort inputs variables being subject to measurement error, that potentially can be different by gender. Since some of the variables of effort are based on students' perceptions, these results might be explained by the fact that male students develop a mistaken perception of their own behaviour and overestimate their effort. Second, it could just express teachers' biased behaviour in favour of girls who show a specific behaviour, in other words: $E[b_{ij}(e)|\text{girl}] > E[b_{ij}(e)|\text{boy}]$. Why this may be the case? It might be related to how effective are the male students in showing or demonstrating their good behaviour to the teacher compared to their female classmates. This explanation is grounded in several studies in psychology (see for example, Salisbury et al. (1999)) that posit that for male students it could not be socially acceptable to be seen interested in school work, because this attitude is in conflict with the society's notions of masculinity. An alternative explanation is related to teachers' expectations of male and female behaviours, either based on gender role attitudes or stereotypes. Teachers may expect students to behave in a certain manner according to specific behavioural patterns, and ignore any behavioural change which diverges from this path. This phenomenon is referred to as *sustaining expectation effect* (Cooper and Good, 1983). Then, if teachers believe that girls present better attitudes to learning or work harder than boys, under the hypothesis of *sustaining expectation effect* teachers may dismiss or ignore any boys' behaviour opposite to this original belief.

TABLE 1.13. Effects of student behaviour on gender grading gaps

	Spanish			Math		
	(1)	(2)	(3)	(1)	(2)	(3)
Female	0.126***	0.112***	-0.058	-0.035***	-0.047***	-0.285***
	(0.002)	(0.002)	(0.037)	(0.002)	(0.002)	(0.033)
Non-blind	0.122***	-1.313***	-1.215***	0.494***	-0.308***	-0.220***
	(0.015)	(0.059)	(0.063)	(0.017)	(0.065)	(0.065)
Non-blind × Female	0.181***	0.147***	-0.047	0.182***	0.144***	-0.031
	(0.004)	(0.004)	(0.049)	(0.004)	(0.004)	(0.046)
Behaviour						
Do homework		0.083***	0.077***		0.096***	0.084***
		(0.002)	(0.003)		(0.002)	(0.003)
Like to study		0.085***	0.082***		0.057***	0.048***
		(0.002)	(0.003)		(0.002)	(0.003)
Grade retention		-0.020***	-0.018***		0.078***	0.086***
		(0.005)	(0.006)		(0.005)	(0.006)
School attendance		0.008***	0.007***		0.014***	0.013***
		(0.000)	(0.000)		(0.000)	(0.000)
Non-blind ×						
Do homework		0.257***	0.257***		0.214***	0.212***
		(0.004)	(0.004)		(0.003)	(0.004)
Like to study		0.090***	0.076***		0.135***	0.120***
		(0.003)	(0.004)		(0.003)	(0.004)
Grade retention		-0.330***	-0.321***		-0.281***	-0.268***
		(0.007)	(0.008)		(0.009)	(0.008)
School attendance		0.013***	0.012***		0.006***	0.006***
		(0.001)	(0.001)		(0.001)	(0.001)
Female ×						
Do homework			0.013***			0.026***
			(0.004)			(0.004)
Like to study			0.007*			0.017***
			(0.004)			(0.003)
Grade retention			-0.006			-0.024***
			(0.007)			(0.007)
School attendance			0.002***			0.002***
			(0.000)			(0.000)
Female × Non-blind ×						
Do homework			0.002			0.005
			(0.005)			(0.005)
Like to study			0.026***			0.029***
			(0.005)			(0.004)
Grade retention			-0.026***			-0.036***
			(0.009)			(0.009)
School attendance			0.002***			0.002***
			(0.001)			(0.000)
R-squared	0.321	0.355	0.355	0.363	0.392	0.393
Observations	1,741,530	1,741,530	1,741,530	1,741,530	1,741,530	1,741,530

Notes: All regressions include student characteristics (father's education, mother's education, household income, indicators of ethnicity and foreign student) and lagged SIMCE test scores, along with their interactions with non-blind test. Standard errors are clustered at school level and are reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

A leading concern is that the behaviour of the students and teachers' grading might be jointly determined, and therefore the estimates in Table 1.13 are biased. To address the problem of simultaneity bias – caused by student's behaviour and teacher's grading influence one another – I use lagged values for behavioural variables (i.e. *Do homework* and *Like to study*). Table A.10 in Appendix A.5 shows that the key findings are robust to these changes.[22]

Overall, the findings in this section indicate that the gender gap in grading could be attributed to students' behaviour, particularly in relation to their academic effort and attitude toward learning. Female students present higher (more positive) values for behavioural variables than male students. It is important to bear in mind that this result may be explained either by effort inputs being more valuable for girls than boys or by teachers' biased behaviour in rewarding that effort.

1.7 Conclusions

Using a precise and comprehensive student-level data on school grades and SIMCE test scores, I find that boys tend to receive lower school grades than girls compared to their SIMCE test scores. Therefore, and in line with the previous literature, there is a gender gap in grading in favour of girls. Moreover, the gender grading gap widens throughout primary school years, and decreases in secondary school.

Gender grading gaps have been utilised by many researchers as a measure of teachers' biased behaviour. This chapter questions the validity of this measure. I show that deviations between non-blind and blind tests might be capturing differences in the tests' production functions for girls and boys, rather than actual teachers' gender biases. I provide evidence that teachers' grading behaviour is not persistent across classes, and that most

[22] Table A.11 in Appendix A.5 reports the results using a 4-point scale to measure students' behaviour (i.e. *Do homework* and *Like to study*), where 1 denotes “I entirely disagree” and 4 denotes “I fully agree”. The results are consistent with those presented in Table 1.13.

of the variation of the gender bias at class-subject level is driven by the characteristics of the class, and not from the teachers' identity. Taken together, these results challenge the idea that teachers grading bias is a fixed characteristic of the teachers. Nevertheless, these results are perfectly compatible with a model which allows teacher's grading to vary with student's behaviour.

By exploiting rich survey information and administrative data on students' school effort, I provide evidence on the mechanisms that could explain the grading gap against boys. I show that the gender gap in grading is not driven by a compositional effect or differences in the structure of the production function of school grades and the SIMCE test scores. Instead, I find that the grading gaps are attributed either by teachers rewarding differently students' behaviour based on their gender, or by female students being more effective in transforming this behaviour into higher school grades.

From a policy perspective, the grading gaps between teachers' assessments and national examinations identified in this chapter have important effects on students' academic achievement for at least two reasons. First, the process of grading plays a crucial role in giving information and feedback to students about their work and performance. Therefore, school grades constitute a powerful channel to transmit information about student ability, allowing students to determine their optimal allocation of effort. Gender gaps in grading compromise the extent to which school grades inform about children's skills and learning. Second, in Chile – as in many other countries – school grades form part of the criteria for admission to higher education institutions. Then, gender gaps in grading have a direct impact on student's probability of admission. In the context of the public debate, the evidence of gender gaps in grading encourages a discussion on the necessity of revising schools grading procedures and policies.

Chapter 2

Let’s stay together: the effects of repeated student-teacher matches on academic achievement

2.1 Introduction

Each year, school managers must allocate teachers to groups of students. Consider a school with two maths teachers, and two groups of students who progress from grade 7 to grade 8. Each teacher could specialise in a particular grade: teacher 1 takes both groups in grade 7, and teacher 2 takes both groups in grade 8. Under this allocation, all students are matched with a new teacher in grade 8. An alternative arrangement is to repeat the student-teacher match, which is called “looping” in the educational literature. Under this allocation, each teacher is assigned to a single group of students which they teach in both grade 7 and 8. Students who remain in the same group between grades will be matched with the same teacher in both grades. Students who change group between grade 7 and 8 will be matched with a new teacher, but will typically still be in a group in which most students have the same teacher in both grades. Does looping have any impact on student achievement? If yes, how and through which mechanisms? This chapter attempts to provide answers to these questions.

Understanding the effect of looping is important for at least two fundamental reasons. First, it is widely used in some school systems. Although systematic quantitative evidence on the prevalence of looping does not appear to be available, it seems to be widespread in German elementary schools (Zahorik and Dichanz, 1994), in Chinese schools at all levels (Liu, 1997) as well as in Finland, Japan, Sweden, Israel and Italy (Tourigny et al., 2019). In the case we study, Chile, over 50% of students progressing from year 7 to 8 have the same teacher in both grades. Thus, measuring the effect of looping-based teacher-student allocations on student outcomes is potentially of great importance. Second, repeating student-teacher matches only requires a re-assignment of existing teaching resources without significant additional costs. Thus, if it works, looping can be a budget-neutral way to improve student achievement.

In this study, we use rich, comprehensive student-teacher data to explore the effect of repeating the student-teacher match on students' test scores for 8th graders in Chile. Unusually, we have information on all student-teacher matches across multiple subjects and multiple years, and we have a national, anonymous measure of student test scores which is uncontaminated by any teacher or school biases in grading. However, even with these data, estimating the causal effect of repeating the student-teacher match is challenging for two reasons. First, because student-teacher matches are non-randomly selected: student-teacher matches which are successful in one year may be more likely to be repeated; certain kinds of teachers may be chosen to teach a specific year group; particular groups of students may be chosen to continue with the same teacher, and so on. Second, even if one could randomly allocate repeat matches, those matches will tend to have more experienced teachers. This arises because, in order to repeat a match, the teacher must have taught at the same school in the previous year, while new matches are drawn from a pool which includes teachers who are recently hired. To deal with these concerns, we estimate the effect of repeated student-teacher matches using plausibly exogenous variation in teacher-student allocation, and we make within-teacher comparisons to control for the resulting experience gap.

We start by exploiting within-school, within-student and within teacher-year variation to control for many of the possible biases which might arise from selection into repeated matches. Repeating a match increases student performance by about 0.02 standard deviations.[1] This is equivalent to the effect of improving teacher quality by 0.1–0.2 standard deviations.[2] A value-added specification yields similar results.

However, fixed effects and value-added methods do not fully mitigate the concern that school managers (or teachers) might decide to repeat matches based on the performance of existing matches. To solve this problem, we consider a situation in which the teacher-student match is broken for exogenous reasons, namely the discontinuity in repeat matches which occurs when teachers reach the legal retirement age (LRA). Effectively, we compare the performance of grade 8 students whose grade 7 teacher reached the LRA in the previous year with grade 8 students whose grade 7 teacher reaches the LRA in the current year. Grade 8 students whose grade 7 teacher reached the LRA in the previous year are far more likely to be allocated a new teacher, and hence are far less likely to experience a repeat match. The discontinuity arises because of small differences in the date of birth of different grade 7 teachers. However, this discontinuity mechanically introduces a difference in teacher experience between repeat and non-repeat classes. To deal with this, we include teacher-by-year fixed effects which remove any variation in experience. Using this discontinuity design, we obtain larger estimates of the benefit of repeating student-teacher matches, of the order of 0.07–0.1 standard deviations.

It seems possible that the benefit of individual repeat matches may not simply aggregate. For example, the positive effects we observe at the student-subject level may be simply due to substitution of a fixed amount of effort by each student towards subjects with familiar teachers, at the expense of subjects with new teachers. We therefore test whether the positive effects of repeat matches aggregate up to the student, class, and school level. At these more aggregated levels a discontinuity approach is not possible and

[1] Hill and Jones (2018) use a similar method in the context of elementary public schools in North-Carolina and obtain similar estimates.

[2] Using estimates from Rivkin et al. (2005) and Rockoff (2004).

so we rely on fixed effects methods which account for possible selection at the class, school-year and subject-level. Reassuringly, we find student, class and school-level estimates are all slightly larger than the equivalent student-subject level estimates.

Finally, we explore several potential channels through which looping may improve student outcomes. Using evidence from a survey of teachers, we assess the effect of repeat matches on the learning environment at the class level. Educational research has emphasised the positive relationship between school effectiveness and a co-operative school environment. The school climate reflects the quality of the relations between the members of the educational community. The literature has shown that a positive and sustained school climate is correlated with higher levels of students' motivation and engagement, school attendance, graduation rates and teacher retention (Thapa et al., 2013). In addition, recent studies (Bryk et al., 2010; Kraft et al., 2016; Klugman, 2017) have established a positive causal impact of school climate on students' achievement on standardised test scores.[3] We find that that in classes with more student-teacher matches, students have higher attendance, teachers report better classroom behaviour and have higher expectations of their students' academic potential.

Very few papers have attempted to formally evaluate the effectiveness of repeat matches for student achievement. An exception is Hill and Jones (2018), who assess the impact of repeat matches on the academic achievement in elementary public schools from North Carolina using a battery of fixed effects.[4] The effect on test scores is positive, significant, and similar to our estimates. We build on their findings by exploiting data on the universe of Chilean students and teachers over a longer period and presenting, to the best of our knowledge, the first estimate of the causal effect of looping utilizing an exogenous variation in teacher-student allocation which allows for student-subject level selection into repeat matches.

Repeating student-teacher matches necessarily implies greater student-teacher familiarity. In this sense, our analysis is related to Fryer (2018), who inves-

[3] For a comprehensive review on school climate literature, see Thapa et al. (2013).

[4] In contrast to our setting, looping is not common in Hill and Jones's case.

tigates the effect of teacher specialisation by subject, and finds that specialisation decreases students' achievement and attendance, and increases student behaviour problems. Fryer suggests that these findings could be explained by the decrease in interactions between teachers and students, caused by teachers' subject specialisation. Our findings support this view in a different context, from a different policy, and provides complementary evidence on how student-teacher familiarity manifests in better classroom behaviour.

A recent literature emphasises complementarities between teacher and student characteristics (e.g. Aucejo et al., 2018; Graham et al., 2020). This implies that improving teaching-to-classroom assignments may lead to better student outcomes. Graham et al. (2020) experiment with different assignments to show that overall achievement in elementary schools in the US can increase by at around 0.02 standard deviations without changes in existing teaching resources. Of course, a precise performance-improving assignment of teachers to classrooms requires information that it is not necessarily available for school managers. Our study complements these findings by providing a simple and feasible assignment rule that delivers results which are at least as large, if not larger.

A number of qualitative and small-scale quantitative studies in the educational literature have investigated the effectiveness of looping, including Bogart (2002), Nichols and Nichols (2002), Cistone and Shneyderman (2004), Tucker (2006) and Franz et al. (2010). Cistone and Shneyderman note that looping is widespread in primary schools in certain countries, including Germany and Japan, but rarely used in others. Most of these studies consider elementary schools: Kerr (2002) stresses that very few studies consider effects on older children. These studies overwhelmingly argue that looping improves student outcomes. For example, Cistone and Shneyderman (2004) find that looping improved student attendance and increased the rate at which students progressed successfully to the next grade. It is commonly suggested that looping has these positive benefits because it saves considerable time at the start of the new school year. Cistone and Shneyderman (2004) argue that looping "allows teachers to save time at the beginning of the second year of the loop by making unnecessary the usual transitional period typically spent on getting acquainted with new students as well as setting classroom rules, expectations, and standards." The same idea is also argued by Burke (1996), Little and Dacus (1999) and Black (2000). A teacher cited by Little and Dacus (1999, p.43) explains: "Gone were the lectures about daily procedures and classroom rules. Gone were the weeks of testing, trying to determine a student's reading level. The teachers and students started the year with a bang and ended further along than the teachers had anticipated." The literature also argues that looping allows teachers to build closer relationships with the students and parents, along with a better understanding of the strengths, weaknesses, and personalities of their students. Looping also allows teachers to implement a smooth transition across grade levels and develop a more cohesive curriculum. This educational literature provides useful insights on how looping may affect the learning process, but does not provide a systematic assessment of its overall causal effect. Our study is a contribution in that direction.

We recognise that looping may also have disadvantages. First teachers may find it more difficult to teach a multi-year rather than single-year curriculum. Second, teachers may lose grade-specific human capital, which Ost (2014) finds contributes up to one-third as much as general teaching experience, at least for maths scores. Finally, even if repeated matches are more efficient, they may also increase inequality in student outcomes, because, as noted by Bogart (2002), some unlucky students will spend two or more years with an ineffective teacher. Assigning students to new teachers each year mitigates these inequality concerns.

The remainder of the chapter is organised as follows. Section 2.2 describes our data and the relevant institutional features of the Chilean school system. Section 2.3 explains the econometric framework and estimates the effect of repeated student-teacher matches at the student-subject level. We begin with fixed-effects methods which maintain the assumption that selection into repeat-matches is exogenous to the quality of existing matches. We then relax this assumption by exploiting the discontinuity at the LRA as a source of exogenous variation in repeat match formation. In Section 2.4 we estimate the effects of repeated matches at the student, class, and school level, which may be more informative as to the effectiveness of a policy of repeating student-teacher matches, since there may be spillover or substitution effects within and between students. In Section 2.5 we report the results from large-scale teacher survey results which support the hypothesis that repeated matches improve behaviour in the classroom and raise teacher expectations of future student performance.

2.2 Data and institutional background

We use three different datasets provided by the Chilean Ministry of Education. First, we use the complete school enrolment records of all students in Chile from 2002 onwards. The database contains yearly information on the students enrolled in primary school (grade 1 to grade 8) and high school (grade 9 to grade 12). These records contain a consistent student ID, a school ID and a "class" ID. In Chilean schools, a class is a fixed group of students who take subjects together: every student in our sample is in the same group (class) in grade 8 for all four subjects we consider. The enrolment records include individual school grades (awarded by teachers) in each subject and the individual attendance rate. The grading system in Chile is 1 to 7 by increments of 0.1, and schools are free to set their own grading standards. To make school grades comparable, we standardise school grades at the school level.[5]

Second, we use comprehensive teachers' administrative records. These records contain information on teacher gender, age, and experience. This database includes the same class ID as in the enrolment records, which allows us to associate each class of students in each subject with a teacher in each year. The enrolment records matched to the teacher records allow us to measure whether a student has the same teacher in a subject for successive years.

Third, we use data on students' achievement in *Sistema de Medición de la Calidad de la Educación* (SIMCE) tests. This is a standardised test administered by the Ministry of Education to all students in certain grades, and

[5] We do not use these school grades as an outcome measure because they may reflect teacher biases as well as student performance (see Chapter 1).

is the main instrument to measure the quality of education in Chile. The SIMCE is administrated by external examiners, and provides information about students' performance relative to the country's National Curriculum Framework. We use standardised test scores for 8th graders in four years: 2004, 2007, 2009 and 2011, in four different subjects: Spanish, maths, social sciences and natural sciences.[6] In these three years, SIMCE tests were taken by $1,056,458$ students, 97.8% of the students enrolled in 8th grade, covering 98.4% of schools in operation.[7]

The SIMCE data also contains information on school characteristics (including whether a school is public or private) and information from surveys of parents and teachers. The parents' survey provides information on family socio-economic background, including mother's schooling and monthly household income (banded). For years 2009 and 2011, the teachers' survey provides information about perception of classroom behaviour and the future performance of the class. Teachers complete a separate survey for each class they teach.

We therefore have information on students $i = 1 \dots N$ who are observed in 8th grade in one of four different years ($t = 2004, 2007, 2009, 2011$). Each student has SIMCE test scores in four subjects $s = 1, 2, 3, 4$. Students are grouped together in classes c. A class-subject combination has a specific teacher j, school k and year t. We start with a sample of $789,270$ students. After excluding observations without valid test scores, student or teacher characteristics, we are left with a sample of $696,482$ students, $46,256$ teachers, $31,837$ classes and $6,260$ schools. Overall, the estimation sample represents 76.3% of the students enrolled in 8th grade who took all the SIMCE tests. Information from teachers about classroom behaviour and future class performance is available for $9,498$ classes for each of the four subjects.

[6] We focus on grade 8 in these four years because we have information on all four subjects' SIMCE test scores, and we exploit the variation across subjects.

[7] The SIMCE test is not taken by students in special education or adult education. In addition, there are cases in which the test cannot be taken because schools are closed temporarily or because students cannot attend. Cuesta et al. (2020) find that high-performing students are more likely to take the SIMCE test, and that the size of this effect varies across school. Our findings, however are based on a within-student design.

A *repeat match* takes place when a student has the same teacher in the same subject as in the previous academic year. We do not consider repeat matches to occur if a student has the same teacher in consecutive years, but not in the same subject. We also do not consider repeat matches to occur if a student returns to the same teacher after a gap.[8]

Students may repeat a grade due to academic failure. Grade retention depends on the students' performance during the school year, as well as their attendance rate. The most prevalent condition for grade retention between grades 4 and 8 is to fail (score below 4.0) in one subject and having a Grade Point Average (GPA) across all subjects lower than 4.45. Students must also attend at least 85% of classes. Grade retention is rare: about 1.8% of the students in grade 8 are repeating the grade. We do not exclude grade repeaters from our analysis because we implement a within-student comparison, as explained in Section 2.3.

Table 2.1 presents descriptive statistics. Panel (a) shows that the outcome (SIMCE test score) and treatment (repeated match) are measured at the student-subject level in grade 8. Repeat matches are common in the 8th grade of Chilean schools.[9] In the estimation sample, 58% of the observations have a repeat match. Panel (a) also shows that repeat matches are less common between grades 6 and 7 (41%) than between grades 7 and 8.[10]

There are no substantial differences in the frequency of repeat matches by subject, shown in panel (b). Because each student has probability of a repeat match of 0.58 in each subject, 8th graders can expect to have a repeat teacher in 2.32 of their four subjects. For each student we also observe sex,

[8] Both are infrequent cases. In the sample, 88.9% of the total matches occur in the same subject. On the other hand, 2.8% of the student-teacher matches in 8th grade present 1 year of gap.

[9] Grade 8 is the final year of primary education, and students will typically move to a different school and have different teachers in grade 9. Students typically remain in the same school between grades 5 and 8, and therefore repeated student-teacher interactions will be common in grades 6, 7 and 8. Our analysis focuses on grade 8 because of the availability of the SIMCE test score information.

[10] We cannot identify repeat matches between grades 5 and 6 for the entire sample because we do not have enrolment data for 2001.

Table 2.1. Descriptive statistics

	Mean	Standard deviation
(a) Student-subject level i, s (N= 2,785,928)		
SIMCE test score	0.00	1.00
1=Repeat match grade 8	0.58	0.49
1=Repeat match grade 6-7	0.41	0.49
(b) Student level i (N= 696,482)		
1=Repeat match (Spanish)	0.57	0.50
1=Repeat match (Mathematics)	0.59	0.49
1=Repeat match (Natural Sciences)	0.59	0.49
1=Repeat match (Social Sciences)	0.58	0.49
Number of repeat matches	2.32	1.30
1=Female	0.51	0.50
Mother's schooling (years)	10.95	3.75
Household's monthly income (000s of CLP)	376.02	468.90
Past GPA	0.09	0.95
Past attendance rate (%)	94.40	5.81
Class size	26.68	8.47
(c) Teacher level j (N= 46,256)		
1=Female	0.68	0.47
Experience (average)	16.34	12.53
Age (average)	43.59	11.80
(d) School level k (N= 6,260)		
1=Public	0.50	0.50
1=Voucher	0.42	0.49
1=Private	0.07	0.26
1=SES 1 (Low)	0.25	0.43
1=SES 2 (Middle-low)	0.33	0.47
1=SES 3 (Middle)	0.23	0.42
1=SES 4 (Middle-high)	0.12	0.33
1=SES 5 (High)	0.07	0.25
1=Urban	0.73	0.44
School enrolment (average)	436.90	402.15
Number of teachers (average)	19.30	14.17
(e) Class-subject level c, s (N= 37,992)		
1=Problems to start the class	0.34	0.47
1=Classroom disruption	0.44	0.50
1=High teacher expectation	0.55	0.50

Notes: Sample comprises students in 8th grade in 2004, 2007, 2009 and 2011 who have valid test scores and a complete set of information on characteristics. Household monthly income is imputed from the mid-point of 15 income bands with widths of 100,000 CLP or 200,000 CLP. The class-subject information in panel (e) is only available for a subset of $9,498$ classes out of $31,837$ classes in total.

family background, past GPA, past attendance rate and class size in grade 8.

In panel (c) we report information at the teacher level, which includes sex, age, and experience. Teachers' experience and age correspond to the average across the four years.[11]

In panel (d) we report information at the school level including size according to enrolment and number of teachers. Schools in Chile may be one of three types: public, private but supported by vouchers and unsupported private.[12] Schools are classified by the Ministry of Education according to the socio-economic status (SES) of their students, based on four variables: father's level of education, mother's level of education, monthly family income and a vulnerability index of the students. The variable ranges between 1 and 5, 5 being indicative of the wealthiest students. Finally, in panel (e) we show information from the SIMCE survey about teachers' perceptions of classroom behaviour[13] and their expectations of their students in the future.[14]

In Table 2.2 we show how the characteristics of the treatment and control groups differ. The raw difference in test score is very small, but repeat matches are positively associated with several factors correlated with *worse* academic performance, including lower family income and lower previous test scores.

[11] In the estimation sample teachers are observed a different number of times across the four years: 52% (24,271 teachers) are observed once; 24% (11,276 teachers) are observed twice; 14% (6,558 teachers) are observed three times, and 9% (4,151 teachers) are observed four times.

[12] For a detailed description of the Chilean school system and education providers, see Santiago et al. (2017).

[13] Teachers were asked about how much they agree or disagree with the following statements: "In this class, it is very hard to start the class lessons" and "In this class, the lessons are often interrupted because I must silence or scold students". The rating scale is "I fully agree", "I agree", "Disagree", "I entirely disagree". Both variables are coded as dummy variables, taking value of one if the teacher answers "I fully agree" or I agree", and zero otherwise.

[14] Teachers were asked "What do you think will be the highest level of education that most students in this class will achieve in the future?". The variable is coded as a dummy variable, taking value of one if the teacher expects that the majority of the class will complete higher education studies and zero otherwise.

Panel (a) shows that repeat matches in grade 8 are themselves correlated with repeat matches in grade 7, which may reflect differences at the school-level in terms of policy towards repeated matches. However, the distribution of repeat matches does not suggest that looping is primarily a school-level policy. Two-thirds of students have variation in repeat matches across subjects (which by definition are taken within the same school). In Appendix B.1 we show that only 15% of the variation in the proportion of repeat matches at the school-subject-grade-year level is accounted for by school fixed effects, and also that very few schools always (or never) use repeat matches.

Panel (b) shows that students who have repeated matches come from lower-income families with less-educated mothers. Repeat matches are positively selected on those measures of academic effort and achievement which are observable by the teacher: past GPA and past attendance rate are both higher for repeat matches. However, repeat matches are *not* positively selected on the anonymised SIMCE test score.[15]

Panel (c) of Table 2.2 shows that repeat matches are significantly more common in public schools, in low socio-economic status schools and in rural schools. There are also important differences in terms of school size and structure, some of which are mechanically related to the probability of repeat matches. Students in smaller schools in terms of enrolment, number of classes, number of teachers and number of teachers per subject are all more likely to have repeat matches. Holding other factors constant, a reduction in the number of teachers who are available to teach a particular subject will increase the probability of repeat matches.

Panel (d) shows that repeat matches have significantly older and more experienced teachers. Repeat matches have teachers with three more years of experience than new matches in 7th grade (i.e. before the current match).

[15] The SIMCE test is taken every year in 4th grade, from 2005 onwards. Therefore, past SIMCE test scores are only available in 2009 (4th grade in year 2005) and 2011 (4th grade in year 2007). 4th grade SIMCE scores are only available for three of the four subjects (Spanish, maths, and natural sciences). As with current SIMCE test scores, scores in 4th grade are standardised to have mean zero and unit variance.

Table 2.2. Characteristics of treatment and control groups

	Treatment group	Control group	Difference	Std. err.
SIMCE test score	0.001	-0.002	0.003***	(0.001)
(a) Previous repeat matches				
1=Repeat match grade 6-7	0.47	0.32	0.154***	(0.001)
(b) Student characteristics				
1=Female	0.51	0.50	0.001	(0.001)
Mother's schooling (years)	10.74	11.26	-0.521***	(0.005)
Household's monthly income	342.37	422.66	-80.287***	(0.567)
Past GPA	0.11	0.06	0.050***	(0.001)
Past attendance rate (%)	94.62	94.09	0.536***	(0.007)
Past SIMCE test score	0.15	0.21	-0.058***	(0.002)
Class size	26.94	26.33	0.613***	(0.010)
(c) School characteristics				
1=Public	0.55	0.44	0.110***	(0.001)
1=Voucher	0.41	0.49	-0.079***	(0.001)
1=Private	0.05	0.08	-0.032***	(0.000)
1=SES 1 (Low)	0.11	0.09	0.027***	(0.000)
1=SES 2 (Middle-low)	0.34	0.30	0.044***	(0.001)
1=SES 3 (Middle)	0.35	0.35	-0.000	(0.001)
1=SES 4 (Middle-high)	0.15	0.19	-0.037***	(0.000)
1=SES 5 (High)	0.05	0.08	-0.033***	(0.000)
1=Urban	0.88	0.91	-0.035***	(0.000)
School enrolment	698.74	820.18	-121.432***	(0.741)
Number of classes	20.09	23.34	-3.248***	(0.018)
Number of teachers	26.29	31.01	-4.722***	(0.023)
Number of subject-teachers	2.66	3.20	-0.542***	(0.002)
(d) Teacher characteristics				
1=Female	0.69	0.68	0.011***	(0.001)
Experience in 7th grade	20.06	16.93	3.124***	(0.015)
Experience in 8th grade	21.06	15.37	5.694***	(0.014)
Δ Experience	1.00	-1.57	2.570***	(0.012)
Age	47.51	42.55	4.962***	(0.013)
Observations	1,618,387	1,167,541		

Notes: The past SIMCE test score is the SIMCE score from grade 4, and is based on $338,941$ and $440,192$ observations in the control and treatment groups respectively. All comparisons are at the student-subject level. The number of subject-teachers is based on the number of teachers in the school between 5th grade and 8th grade, because the majority of the teachers from the first cycle (grades 1–4) are general teachers, and they teach all the main subjects to a particular class. In the case of the four years analysed (2004, 2007, 2009, 2011), 95% of the teachers from the first cycle teach more than one subject. In contrast, 44% of the teachers from 5th grade to 8th grade are subject specialist, and teach only one subject. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Repeat matches have teachers with *six* more years of experience than new matches in 8th grade. More experienced teachers are more likely to get repeat matches, and, by definition, repeat matches have a teacher with one more year of experience than in the previous year. In contrast, new matches draw a new teacher who has more than two years *less* experience than their teacher in the previous year. This arises because, by definition, teachers who have repeat matches in 8th grade must have worked at the school in 7th grade, whereas new matches may draw a teacher who is new to the school.

Given these differences in students, schools and teachers between repeat matches and new matches, it is important to note that we observe the same student (by definition in the same school) in multiple subjects, some of which are repeat matches and some of which are new matches, and we observe the same teacher with multiple classes,[16] some of which are repeat matches and some of which are new matches. This enables us to control both for unobserved fixed student effects and unobserved fixed teacher effects, which greatly reduces any concerns about selection on the basis of these characteristics.

2.3 The effect of repeat matches at the student-subject level

As shown in Table 2.2, a simple comparison of repeat matches and new matches may be misleading because repeat matches are not randomly assigned: repeat matches have systematically different students, teachers, and schools. These differences may arise because of teacher and student sorting within schools, and because of teacher and student mobility between schools. Previous research has established the existence of teacher sorting within schools: less-experienced, minority and female teachers are systematically sorted to lower-performing students (Clotfelter et al., 2005, 2006; Feng, 2010; Kalogrides et al., 2013). Moreover, qualitative research shows that school leaders base their staffing decisions on a combination of teach-

[16] A small fraction of teachers are observed in more than one school.

ers' performance (measured by their students' test scores) and teachers' preferences (Cohen-Vogel, 2011; Kalogrides et al., 2013; Osborne-Lampkin and Cohen-Vogel, 2014). Teacher and student mobility between schools may also cause differences in the proportion of repeat matches, and it seems likely that the decision to move schools will not be exogenous with respect to student outcomes.

Our data allow us to control for differences in fixed student characteristics by using the within-student variation across subjects, taking advantage of the fact that we observe students' test scores in four different subjects.[17] In addition, since students attend the same school and the same class for all subjects, student fixed-effects will also control for selection bias as a result of differences in school or class characteristics. The inclusion of student fixed effects also addresses two specific sources of selection bias: parental choice of school and grade retention. First, parents' decision whether to move their child to another school could lead to a selection issue if parents take this decision based on, for instance, how well their children are matched with their teachers in a particular school. In the estimation sample 7.8% of the students change school between grade 7 and grade 8. Second, students who repeat the grade due to academic poor performance are significantly less likely to have a repeat match. In the estimation sample, about 1.8% of the students are grade repeaters, of which 65.7% do not have the same teacher again. Grade repeaters are more likely to come from low-income families, to have less educated mothers, and to have lower test scores. The inclusion of student fixed effects deal with both these potential biases, since children attend the same school for all subjects, and grade repeaters re-take all subjects.

As well as addressing selection bias, the inclusion of student fixed-effects allows us to estimate the effectiveness of repeat-matches independent of any effect of a group of students staying together between grades. It seems possible that student-student familiarity (in addition to student-teacher familiarity) has a causal effect on student outcomes, and the process of as-

[17] Many cross-sectional studies exploit within-student variation to identify effects of teacher characteristics and teaching practices (Dee, 2007; Clotfelter et al., 2010; Bietenbeck, 2014; Bietenbeck et al., 2018; Paredes, 2014; Lavy, 2015; Comi et al., 2017).

signing the same teacher to a group of children necessarily implies that the group (or at least the majority of the group) stay together between grades. The fixed-effect strategy we use compares the same student across subjects in the same year, and this student will have the same classmates for all subjects, so we are effectively comparing outcomes for the same group of students, some of whom have a repeat match and some of whom do not.

Our method also allows us to control for differences in fixed teacher characteristics by using the within-teacher variation across classes, taking advantage of the fact that we observe the same teacher in several classes. Further, and in contrast to students, we observe the same teacher in multiple classes at four different points in time (2004, 2007, 2009 and 2011) which allows for the inclusion of teacher-by-year fixed effects. As was clear from Table 2.2, there is inevitably a strong relationship between repeating the student-teacher match and teacher experience. Even if repeat-match teachers were drawn randomly, these teachers by definition must have worked in the same school at $t-1$, but new match teachers are drawn from the pool of available teachers which includes those who are new to the school. In addition, repeat-match teachers are not drawn randomly: they have about three more years of experience, on average. Thus, an unconditional comparison of classes which have a repeat match with those that do not conflates the advantages of a repeat match with any advantages of having a teacher who has nearly six years more experience (see Panel (d) of Table 2.2). Since experience is fixed for a given teacher in a given year, the inclusion of teacher-by-year fixed effects controls for this large difference in experience.

Thus, our first model to identify the effect of a repeat match is:

$$y_{is} = \beta_1 R_{is} + \mu_i + \mu_s + \mu_{jt} + \epsilon_{is}, \tag{2.1}$$

where y_{is} is the standardised SIMCE test score of student i in grade 8 in subject $s = 1, 2, 3, 4$ (maths, Spanish, social sciences, natural sciences). Each student is observed in grade 8 in one year $t = 2004, 2007, 2009, 2011$, and therefore i identifies t. For a particular student-subject-year combination we observe the identity $j = J(i, s, t)$ of the teacher. In (2.1) each student i appears in only one school in one year, whereas teachers j appear in multiple classes and years and may also be observed in more than one school. R_{is} is an indicator variable which takes the value 1 if there is a repeat match, which occurs if $J(i,s,t-1) = J(i,s,t)$. As discussed, the model includes student, subject and teacher-by-year fixed effects.[18]

Table 2.3 presents estimates of versions of Equation (2.1) with the inclusion of different fixed-effects. Across all specifications, the results show a positive and significant effect of repeating the student-teacher match on student's SIMCE test scores. The raw effect in Column (1) is small, but recall from Table 2.2 that repeated matches are far from randomly assigned, and are often associated with baseline characteristics which themselves are associated with lower test scores. Including student fixed effects in Column (2) increases the effect to 0.026σ, while the inclusion of both student and teacher effects in columns (3) and (4) reduces the effect to 0.017σ. The inclusion of teacher-by-year fixed effects in Column (4) controls for any effect of differential experience between teachers who repeat matches and those who do not and increases the estimate to 0.019σ.[19] We find no evidence that the size of the effect varies across subjects: an F-test of the interactions between R_{is} and μ_s is insignificantly different from zero. It is also possible to replace the teacher-by-year fixed effects μ_{jt} with teacher-by-subject-year fixed effects μ_{jst} to ensure that we are not conflating looping with an effect from non-looping teachers teaching different (possible less preferred) subjects. The inclusion of μ_{jst} slightly reduces the estimate to 0.016σ.

In Column (5), we include as a control lagged test scores at the student-subject level (Rivkin et al., 2005; Harris and Sass, 2011; Chetty et al., 2014*a*). This is a value-added model which controls for within-student differences in ability across subject which may be correlated with the looping decision. However, the SIMCE test score information for these students is only available in grade 4 and grade 8, so this does not deal with the

[18] The model is estimated using the methods developed by Correia (2016) and Guimaraes and Portugal (2010).

[19] Excluding students who have no variation in R across subjects makes almost no difference, with an estimated effect of 0.018σ (0.002).

Table 2.3. Effect of repeat student-teacher match on test scores: fixed-effect estimates

	(1)	(2)	(3)	(4)	(5)	(6)
Repeat match grade 7-8	0.003**	0.026***	0.017***	0.019***	0.021***	0.020***
$R_{is} = 1$	(0.002)	(0.001)	(0.001)	(0.002)	(0.004)	(0.004)
SIMCE score in grade 4					0.276***	0.276***
					(0.002)	(0.002)
$R_{is} = 1$ grade 6-7						0.014***
						(0.003)
$R_{is} = 1$ grade 5-6						0.007***
						(0.002)
$R_{is} = 1$ grade 4-5						0.006
						(0.004)
Subject FE	Yes	Yes	Yes	Yes	Yes	Yes
Student FE		Yes	Yes	Yes	Yes	Yes
Teacher FE			Yes			
Teacher FE × Year FE				Yes	Yes	Yes
R-squared	0.000	0.793	0.808	0.812	0.849	0.849
Observations	2,785,928	2,785,928	2,785,928	2,785,928	759,597	759,597

Notes: Dependent variable is the student's SIMCE test score in grade 8. In all columns, treatment is the student-subject measure of repeated match R_{is} in grade 8. Standard errors are clustered at the student level. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

problem that the decision to loop may be based on match quality in grade 7. The sample in Column (5) is significantly smaller because the grade 4 SIMCE score is only available in 2009 and 2011, and only in three of the four subjects.[20] The inclusion of lagged SIMCE scores makes almost no difference to the estimate. Finally in Column (6) we deal with the concern that repeat matches may be correlated with earlier looping decisions by including as controls the value of R_{is} in grades 5, 6 and 7. Once again, this makes almost no difference to our estimate of the effect of looping on test scores in grade 8.

Our estimates are very similar to those reported by Hill and Jones (2018) for younger students' maths scores in North Carolina elementary schools (grades 3–5) using a similar specification, but which also includes lagged

[20] Repeating the Column (4) model on this reduced sample yields an estimate of 0.020σ (0.004).

test scores as a control variable.[21] The outcome measure used by Hill and Jones (2018) is a maths score which was reported by the teacher themselves, rather than an anonymised national test score as in our case. This suggests that the use of an anonymised test score, as in our case, is not crucial for finding positive effects from repeated matches.

The remaining source of variation in (2.1) is the error term ϵ_{is}, which varies at the student-subject (equivalent to the student-teacher) level. If repeat matches are formed non-randomly with respect to this "match quality" term, then estimates of β_1 will still be biased even after controlling for student and teacher fixed-effects. Schools or parents may both make decisions about which class-teacher matches to keep together in grade 8 on the basis of their performance in grade 7. As a result, class-teacher matches are endogenously destroyed, and the effect of a repeat match will be confounded by survivor bias.

Unfortunately, we do not have the SIMCE test score in grade 7 for these students. However, we can use information on SIMCE scores in grade 6 to predict match formation in grade 7. To do this we estimate (2.1) on a sample of all grade 6 students for whom we have SIMCE test scores[22] and calculate $\hat{\epsilon}_{is,6}$, the residual for each student-subject observation. We then calculate, for each student-subject observation, the average residual of their classmates, $\bar{\hat{\epsilon}}_{i's,6}$ and estimate whether these residuals have any effect on the formation of repeat matches in grade 7:

$$R_{is,7} = \gamma_1 \hat{\epsilon}_{is,6} + \gamma_2 \bar{\hat{\epsilon}}_{i's,6} + \mu_{jt} + \eta_{is,7}. \tag{2.2}$$

In this model, γ_1 captures whether students whose individual residual is high are more likely to remain with the same teacher in grade 7, while γ_2 captures whether students whose classmates have high residuals are more likely to remain with the same teacher in grade 7. Our estimate of γ_1 is negative, but extremely small and insignificantly different from zero

[21] Hill and Jones report an effect size of 0.018σ (0.005). The increased precision of our estimates likely reflects the much wider prevalence of repeat matches in our data; Hill and Jones report that only 3% of students experience a repeat match in their data.

[22] We have information on SIMCE scores for Spanish and maths in 5 years (2013, 2014, 2015, 2016 and 2018).

(-0.0003 (0.0004)). Our estimate of γ_2 is slightly larger but still insignificantly different from zero at conventional levels (-0.009 (0.005)). Thus, we find no evidence that student-subject combinations which perform better than expected are more likely to lead to repeat matches.

Nevertheless, because we cannot directly control for endogenous selection, we also consider a regression discontinuity approach which exploits the discontinuity in the probability of a repeat match which occurs because of small differences in teachers' date of birth in the year before the grade 8 observation which affect exactly when teachers reach the legal retirement age (LRA). A student whose teacher reaches the LRA in grade 7 is less likely to match in grade 8, because that teacher is more likely to retire. The discontinuity which occurs at the LRA is plausibly exogenous with respect to ϵ_{is}. Clearly, the retirement decision itself is unlikely to be exogenous with respect to student performance, as noted by Fitzpatrick and Lovenheim (2014). Hanushek et al. (2004) also argue that there are teacher selection effects with age which can bias estimates of the returns to teacher experience. However, although match (or teacher) quality may vary with teacher age, there is no reason why they would be discontinuous at the LRA itself. Manipulation of (reported) teacher date of birth is implausible in this setting.

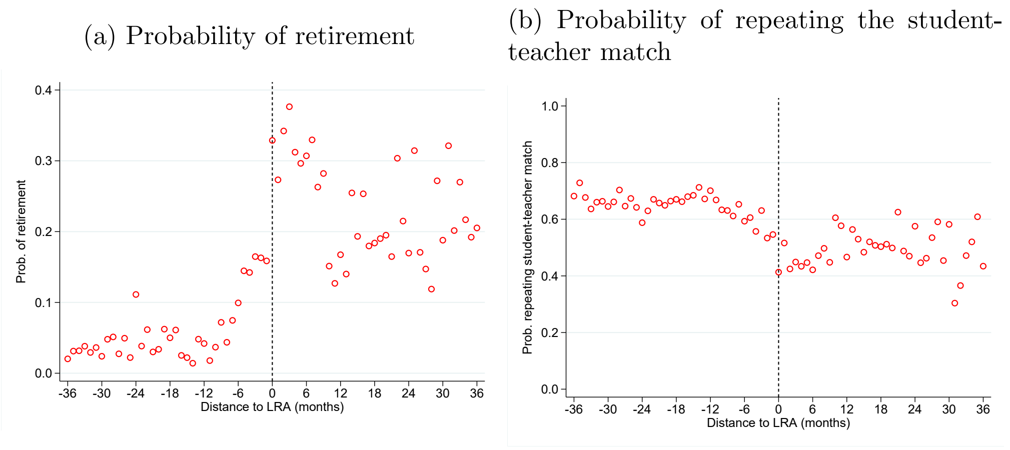
In Chile, the LRA is 65 for men and 60 for women, but teachers are not obliged to retire from the labour market at that age. The law permits early retirement, provided that teachers meet some financial requirements.[23] The school-year starts during the first week of March and finishes in late November or early December. School administrators assign teachers to classes on the assumption that teachers will remain in the school until the end of the school year. Each teacher's exact date of birth is recorded, and using this we calculate age for each teacher on the last day in February in each year (2004, 2007, 2009 and 2011), i.e. the day before the school year starts. Our key identifying claim is that teachers who reach the LRA just before the 1 March are significantly more likely to retire than teachers who reach the

[23] To retire early, workers are required to have sufficient pension resources to fund a replacement rate of 70 percent with respect to their average salary over the previous 10 years, and a minimum pension set by law.

LRA just after 1 March. For example, a grade 7 class in the 2006 school year whose (female) teacher reaches 60 in February 2007 is less likely to have the same teacher in grade 8 than a class whose teacher reaches 60 in March 2007.

Although we do not have a formal measure of retirement, we observe the population of school-teachers in Chile in each year and therefore we can infer retirement quite precisely from the disappearance of a teacher from the data for the next five years. In the left-hand panel of Figure 2.1 we show that the probability of retirement increases quite sharply (but with no discontinuity) for teachers who will reach the LRA in the next school year, and then jumps by over 10 percentage points between teachers who reach the LRA in February (distance to LRA$= 0$) and those who reached it in March (distance to LRA$= -1$). In the right-hand panel of Figure 2.1 we show that this discontinuity is reflected in a sharp 15 percentage point reduction in the probability of a repeat match.

Figure 2.1. Discontinuity in retirement at the LRA and repeat matches, distance in months


(a) Probability of retirement

(b) Probability of repeating the student-teacher match

Notes: A teacher is considered retired if she does not appear in the next five consecutive years in the administrative records of Ministry of Education. The distance to the legal retirement is the difference between the current age and the LRA, recorded in months. The distance to the legal retirement is zero for those teachers whose birthdays are in February and therefore reach the LRA in the last month of the previous school year.

We therefore have a fuzzy-RD design with distance to the LRA of each student-subject combination in grade 7, denoted D_{is}, as the running variable, which can be measured in days. Following Imbens and Lemieux (2008), the RD estimator is defined as:

$$\tau_{RD} = \frac{\lim_{D_{is}\downarrow 0} E[y_{is}|D_{is}=0] - \lim_{D_{is}\uparrow 0} E[y_{is}|D_{is}=0]}{lim_{D_{is}\downarrow 0} E[R_{is}|D_{is}=0] - \lim_{D_{is}\uparrow 0} E[R_{is}|D_{is}=0]} = \frac{\tau_y}{\tau_R} \qquad (2.3)$$

As before, y_{is} denotes the SIMCE test score in 8th grade. The RD estimator corresponds to the ratio between the average intention-to-treat effect (τ_y) and the first-stage effect (τ_R).

We adopt a local polynomial modelling approach to approximate the functional form of τ_y and τ_R. This method uses only the observations that lie between $-h$ and $+h$, where h is a positive bandwidth. Local polynomial estimation involves choosing a kernel function to weight the observation within the interval $[-h, +h]$. We use a triangular kernel function, which gives the maximum weight at $D_{is} = 0$. We use a polynomial of order one, that is to say, we run a local-linear regression within the bandwidth. To select the bandwidth, we follow the procedure proposed by Calonico et al. (2014) by selecting the parameter h that minimises an approximation to the asymptotic mean squared error (MSE) of the point estimator ($\hat{\tau}^{RD}$). Intuitively, choosing a small bandwidth will reduce the approximation bias, but at the same time will increase the variance of the estimated coefficient. For inference, we use robust confidence intervals based on bias-correction following Calonico et al. (2014).

The validity of the discontinuity approach is based on the usual three IV assumptions. First, a relevance condition, that the LRA has a strong effect on the probability of teacher retirement, which in turn affects the probability of repeating the student-teacher match. We have already seen that the discontinuity is a powerful predictor of retirement, and therefore of repeat matches. Second, the instrument exogeneity condition, in this case that the discontinuity at the LRA is exogenous with respect to student potential outcomes. In Figure B.2 in Appendix B.2 we provide evidence that differences in observable characteristics either side of the LRA are very small and almost all insignificantly different from zero compared to the differences in the treated and controls. Figure B.3 shows that density of the running variable shows no sign of manipulation at the cutoff.[24] In order to deal with any remaining imbalance, we supplement our RD estimates with parametric RD estimates which allow for within-student and within-teacher comparisons. Third, we require that the discontinuity effect on student outcomes is *only* driven by its effect on repeat matches. There are two threats to the exclusion restriction. Even if the variation in repeat matches which is caused by the discontinuity is as good as randomly assigned, this variation also causes (quite large) variation in teacher experience. To deal with this, we also consider parametric RD models which allow for the inclusion of teacher-by-year fixed effects which remove any variation in experience between repeated and non-repeated classes.

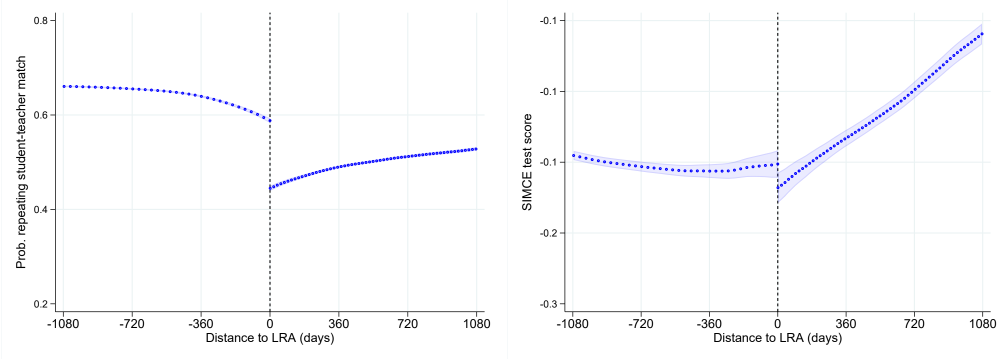
The resulting RD estimates are local for a very specific type of repeat match. The discontinuity will identify the causal effect of a repeat match with an experienced teacher who complies with the discontinuity. In other words, a teacher whose retires at the LRA. If the effect of repeat matches itself varies with teacher experience, then the IV estimates will not be comparable to the fixed-effect estimates from (2.1).

The regression discontinuity results are illustrated in Figure 2.2, which shows the first stage estimate of τ_R in the left-hand panel and the reduced form estimate of τ_y in the right-hand panel. As we anticipated, the first stage shows a large negative effect: students whose teacher reaches the LRA in grade 7 are about 17 percentage points less likely to repeat the match in grade 8. The reduced-form effect on SIMCE test score is about -0.03σ: students whose teacher reaches the LRA in grade 7 have lower test score outcomes in grade 8.

In Appendix B.2 (Figure B.2) we provide some evidence on the exogeneity assumption by estimating the non-parametric RD model but using a wide range of measured characteristics as the outcome variable. For reference, we also show the estimated difference in means from a raw comparison of treated and controls. In the top panel, differences in means are greatly reduced and, in most cases, insignificantly different from zero. However,

[24] The manipulation test of Cattaneo et al. (2018) estimates the density of the running variable either side of the cutoff using a local polynomial and yields a p-value of 0.1314.

FIGURE 2.2. Conditional mean plots by local linear regressions: probability of repeating student-teacher match and SIMCE test score



Notes: Panel a) the probability of repeating the student-teacher match against the distance to the LRA between $[-1,080, 1,080]$ days. Panel b) SIMCE test scores against the distance to the LRA between $[-1,080, 1,080]$ days. The distance to the legal retirement is the difference between the current age and the LRA. The distance to the legal retirement is zero for those teachers whose birthdays are 1st March and reach the LRA in that day. The graphs show conditional mean plots using local linear regression within a MSE-optimal bandwidth (bandwidth = 965 days), with triangle kernel function and a 1st order polynomial, on a grid of 500 points on each side of the cutoff.

some small imbalance remains. One possible explanation for this is that early retirement decisions may also be discontinuous at the February-March threshold, and those decisions may be related to school type.[25] In the bottom panel we repeat the exercise but include controls for school type (public, private, voucher). We now see even less imbalance across the discontinuity. The only exception remaining is household income, which is slightly higher for children whose teacher's age is just below the LRA. As noted, income in the SIMCE data is reported in 15 bands from which we imputed a continuous variable. All of these bands are balanced across the discontinuity once we control for school type, as shown in Table B.2. The possibility that there are small imbalances at the discontinuity motivates us to also consider parametric RD models which allow for within-school and within-teacher comparisons.

Figure 2.2 implies a causal effect of repeat matches which is substantially larger than the fixed-effect estimates in Table 2.3, because the ratio of τ_y

[25] For example, if some schools encourage teachers to retire at the end of the school year *before* they reach the LRA, there may be imbalance in characteristics at that threshold in the following year.

and τ_R is approximately 0.2σ. In Column (1) of Table 2.4 we report a non-parametric RD estimate of 0.158σ which corresponds exactly to Figure 2.2. However, this large estimate may arise because we are conflating the repeat-match effect with an experience effect: although the discontinuity as good as randomly selects students into repeat matches, the discontinuity also selects students into more or less experienced teachers. We test whether this large estimate is due to the experience effect by applying exactly the same RD model to teacher experience. Our estimate of the teacher experience effect of the discontinuity is very large: over 21 years with a standard error of less than one year. This means that, although we can plausibly claim that the LRA discontinuity as good as randomly breaks up student-teacher pairs in grade 7, it has a large causal effect both on the probability of repeating the match *and* on the experience of the teacher in grade 8.

Therefore, in Column (2) of Table 2.4 we adopt a linear functional form for the distance to the LRA, which has a number of advantages. First, it greatly improves estimation precision. Second, and more importantly, it allows us to include student and teacher-by-year fixed-effects, which sweep out any non-random selection of new teachers in comparison to the teachers of continuing matches. In particular, it allows us to control for the experience effect of looping. As expected, this method reduces the effect of looping and produces an estimate of 0.110σ with a substantially smaller standard error. Our estimate of the returns to experience suggests that about half the difference between the results in columns (1) and (2) can be accounted for by the loss of experience which is associated with getting a new teacher in grade 8.[26]

A disadvantage of the simple linear model reported in Column (2) is that Figure 2.2 suggests that the relationship between looping and age is somewhat non-linear in the year before and after the LRA. Therefore, in Column (3) we report a quadratic model which allows for this non-linearity, but which also allows for the inclusion of student and teacher by year fixed-

[26] Our data allows us to estimate the likely effect of this loss of experience since we have a clean measure of student achievement and teacher experience. Following the method of Harris and Sass (2011) our return to experience model in Appendix B.3 predicts that losing a teacher at the LRA with 25 years of experience (the sample mean) and replacing them with a new teacher causes a loss in student test scores of 0.024σ.

Table 2.4. Effect of repeating the student-teacher match on test scores: regression discontinuity results

	Non-parametric (1)	Linear with fixed effects (2)	Quadratic with fixed effects (3)
τ_R (First stage)	-0.137***	-0.121***	-0.114***
	(0.004)	(0.004)	(0.005)
τ_y (Reduced form)	-0.022**	-0.013***	-0.014**
	(0.009)	(0.005)	(0.006)
τ_{RD}	0.158**	0.110***	0.124**
	(0.063)	(0.038)	(0.051)
Student FE		Yes	Yes
Subject FE		Yes	Yes
Teacher FE × Year FE		Yes	Yes
First-stage R-squared		0.873	0.873
First-stage F statistic	941	1,041	566
95% C.I.	[.035 ; .28]		
Effective observations: Left	200,343		
Effective observations: Right	109,731		
Optimal Bandwidth	964.830		
Observations	2,785,928	2,785,928	2,785,928

Notes: Dependent variable is the student's SIMCE test score in grade 8. Treatment is the student-subject measure of repeated match R_{is}. Column (1) presents results based on Calonico et al. (2014) with a polynomial of order one and weighted by a triangular kernel. Column (2) includes distance to the LRA linearly, and the interaction between the distance to the LRA and the indicator variable for reaching the LRA. Column (3) includes a quadratic interaction between distance to the LRA linearly and the indicator variable for reaching the LRA. Standard errors in Column (1) are calculated using Calonico et al. (2014). Standard errors are clustered at the student-level. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

effects. The quadratic model yields an estimate of 0.124σ, with a slightly larger standard error than the linear model.

Do the positive effects of repeated matches occur for every subject? In Appendix B.4 we investigate this issue by estimating the linear RD model separately for Spanish, maths, natural sciences, and social sciences. In these models we cannot control for student fixed effects because each student is observed only once in each subject in grade 8, but we can still control for teacher-by-year fixed effects because teachers take multiple classes in the same subject (both within and across years). In all four subjects there is a strong negative effect of reaching the LRA on the probability of repeating the match. This effect is weaker in Spanish, but very consistent in the other three subjects. The reduced form estimate of τ_y is negative in all four subjects, implying that the estimate of τ_{RD} is positive in all four subjects. However, standard errors are considerably larger than in the equivalent linear model because the sample size is much smaller, so it is hard to make precise statements about the difference in effectiveness across subjects. The effect appears smallest in natural sciences and largest in Spanish, but these results are too imprecise to draw more conclusions about the efficacy of repeat matches in different subjects.

All our RD estimates are larger than the fixed-effects and value-added estimates. This seems unlikely to be the result of strong negative selection into repeat matches. The RD estimates are local in that they relate to very experienced teachers whose retirement decision is affected by reaching the LRA. Therefore, our results suggest that the benefits of looping may be significantly greater for more experienced teachers. However, a natural concern is that, instead, this reflects a failure of the exclusion restriction. Since our parametric models include grade 8 teacher-by-year fixed effects, any failure of the exclusion restriction can only plausibly come from discontinuities in grade 7. A particular concern is that the discontinuity may have an effect on teacher effort in grade 7 which may in turn effect outcomes in grade 8. We test of this restriction by considering a sample of students who change school between grade 7 and grade 8. These students cannot loop,[27] and their grade 8 teacher is selected independently of the grade 7 discontinuity, which leaves grade 7 teacher effort as the only channel by which the discontinuity can affect test scores in grade 8.

Table 2.5 reports estimates of the reduced form τ_y when the sample is restricted to school-movers. All three estimates are insignificantly different from zero, although we note that the non-parametric estimate in Column (1) are imprecise and of the same size as in Table 2.4. More encouragingly, the parametric estimates of τ_y are close to zero. These estimates support our claim that the effect of the LRA discontinuity on test scores operates through its effect on repeat matches.

[27] A tiny number of school movers do in fact have the same teacher in grade 8, presumably because their teacher moved simultaneously or because their teacher had classes in multiple schools.

Table 2.5. Effect of the discontinuity on school-movers: reduced form regression discontinuity results

	Non- parametric (1)	Linear with fixed effects (2)	Quadratic with fixed effects (3)
τ_y	-0.022 (0.029)	-0.0004 (0.010)	-0.010 (0.013)
Student FE		Yes	Yes
Subject FE		Yes	Yes
Teacher FE × Year FE		Yes	Yes
95% C.I.	[-0.079;0.035]		
Effective observations: Left	181,754		
Effective observations: Right	11,706		
Optimal Bandwidth	1280.411		
Observations	193,460	193,460	193,460

Notes: Sample restricted to students who changed school between grade 7 and grade 8. Dependent variable is the student's SIMCE test score in grade 8. Column (1) presents results based on Calonico et al. (2014) with a polynomial of order one and weighted by a triangular kernel. Column (2) includes distance to the LRA linearly, and the interaction between the distance to the LRA and the indicator variable for reaching the LRA. Column (3) includes a quadratic interaction between distance to the LRA linearly and the indicator variable for reaching the LRA. Standard errors in Column (1) are calculated using Calonico et al. (2014). Standard errors are clustered at the student-level. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

2.4 The effect of repeated matches on students, classes, and schools

The comparison we made in Section 2.3 was between individual student-teacher matches that repeat and those that do not. The great advantage of this comparison is that allows us to make within-student and within-teacher comparisons, and our RD strategy also allows to control for endogenous matches at the student-subject level. However, repeat matches may have spillover effects on untreated units. At the student level, a student may allocate greater effort to subjects in which there is a repeat match, but at the same time allocate less effort to non repeat-match subjects. If this was the case, increasing the number of matches at the student level would be less effective. At the class level, if repeat matches allow teachers to save time, there will be benefits to all students in the class, regardless of whether students are individually repeating the match. On the other hand, if repeat matches are beneficial because of greater familiarity between teacher and student, it might not be beneficial for those who join a class in which most other students have a familiar teacher. Indeed, it seems possible that it might actually be harmful if teachers focus their efforts on students with whom they are familiar. At school-level, the allocation of teachers is a joint problem where repeating a match for one teacher has some implication for all other allocations within that school. In this section we therefore aggregate our data and use fixed-effect methods to examine whether the positive effects at the student-subject level carry over to student, class, and school-level.

Our student-level model is:

$$\bar{y}_i = \beta_1 \bar{R}_i + \beta_2 x_i + \mu_c + \epsilon_i, \tag{2.4}$$

where $\bar{y}_i$ is student i's average SIMCE score across all four of their grade 8 subjects, and $\bar{R}_i$ is the proportion of their four subjects in which they have the same teacher as in grade 7. The model includes class fixed-effects μ_c and a set of pre-determined student-level characteristics x_i. The variation we are exploiting here is the within-class variation in repeat matches which arises because not all students in a particular class in grade 8 will have had the same teacher in grade 7.

Our class-subject model is:

$$\bar{y}_{cs} = \beta_1 \bar{R}_{cs} + \beta_2 x_j + \mu_c + \mu_s + \epsilon_{cs}, \tag{2.5}$$

where $\bar{y}_{cs}$ is the average SIMCE score of all students in class c and subject s in grade 8, and $\bar{R}_{cs}$ is the proportion of the class-subject combination who have the same teacher as in grade 7. The model includes class μ_c and subject μ_s fixed-effects and a set of pre-determined teacher-level characteristics x_j. The variation we are exploiting here comes from that the fact that $\bar{R}_{cs}$ varies across subject within class. Note that in both (2.4) and (2.5) there is no time variation because each student and class is observed in only one year.

Finally, our school-subject-level model is:

$$\bar{y}_{kst} = \beta_1 \bar{R}_{kst} + \beta_2 x_{ks} + \mu_{kt} + \mu_s + \epsilon_{ks} \tag{2.6}$$

where $\bar{y}_{kst}$ and $\bar{R}_{kst}$ are the school-subject-year level averages of y_{is} and R_{is} in Equation (2.1); μ_{kt} is a school-by-year fixed effect; μ_s is a subject fixed effect; x_{ks} is a vector of characteristics of the school that vary across subjects and years (specifically, the proportion of female teachers and average experience). The parameter of interest is β_1. Note that at the school level we have four cohorts of grade 8 students from 2004, 2007, 2009 and 2011, and hence (2.6) has time variation. Equation (2.6) relies on variation within schools across subjects and across time for identification. This allows us to rule out selection into schools which might occur if, for example, better schools have more (or less) repeat matches. Also, exploiting the fact that we observe the same school for different cohorts, it is possible to include a school-by-year fixed effect μ_{kt}. This effect will remove all differences between school cohorts which might arise if repeat matches are used for some cohorts and are related to cohort-specific unobservable shocks.

TABLE 2.6. Effect of repeat matches on test scores at student, class and school-level

	Student level ($\bar{y}_i$)	Class-subject level ($\bar{y}_{cs}$)	School-subject-year level ($\bar{y}_{kst}$)
	(1)	(2)	(3)
Proportion of repeat matches	0.039***	0.029***	0.032***
	(0.006)	(0.002)	(0.002)
Class FE	Yes	Yes	
Subject FE		Yes	Yes
School-by-year FE			Yes
Student controls	Yes		
Teacher controls		Yes	Yes
R-squared	0.414	0.916	0.911
Observations	696,482	127,348	82,524

Notes: In each model the independent variable is the proportion of repeat matches at that level. Column (1) includes controls for students' gender, household income, mother's education, and attendance rate in grade 7. Columns (2) and (3) include controls for teachers' gender and experience. Standard errors are clustered at the class-level. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

All three estimates are positive and significant, consistent with a positive effect of repeat matches on students, classes, and schools. It is striking that all three estimates are larger than the comparable student-subject level estimates in Table 2.3. This can partly be explained by the fact that these models do not control for teacher fixed-effects – the exclusion of teacher effects in Column (2) of Table 2.3 produces larger estimated effects at the student-subject level as well. Larger effects are also consistent with positive spillovers from repeat matches within students, classes, and schools.

2.5 Classroom behaviour and teacher expectations

Our results consistently show that repeating the student-teacher match results in a positive effect on student test scores. We find these effects at various different levels of aggregation. In this section, we provide further evidence of the effectiveness of repeat matches on the behaviour of students and the views of their teachers. Specifically, we estimate the effect of repeat matches on student attendance, student behaviour and teacher expectations of their students.

The student enrolment data contains a record of student attendance measured at the student level (we do not observe attendance by subject separately for each student), so we estimate a variant of (2.4) and regress the standardised attendance rate on $\bar{R}_i$, the proportion of subjects in which the student has a repeat match in grade 8. As in (2.4), the model includes class fixed effects and therefore relies on within-class variation.

An independent measure of student behaviour is available from the survey of teachers about their perception of classroom behaviour and the future performance of the class, which is available in 2009 and 2011. Although teachers who complete these surveys are clearly aware of whether their class is a repeat match or not, it is nevertheless a measure which is entirely independent of the anonymised SIMCE test score. Teachers do not know what their students' test scores are, and so this cannot influence their responses to the survey.[28] There are three survey responses of interest. Teachers are asked if they face behavioural problems at the beginning of the class and disruptions during the class. These two outcomes are coded as binary variables, taking value of 1 if they are strongly agree or somewhat agree, and 0 otherwise. In addition, teachers are asked about the level of education that most of the class will achieve. The teacher expectation is coded as a binary variable, taking value of 1 if the teacher expects the majority of the class would finish any type of higher education (either a professional degree or a technical degree) or postgraduate studies. Our data is at the class-subject level, so we use a variant of (2.5) where the dependent variable is our measure of teacher perception (behaviour, expectations) for class c subject s, and the treatment is $\bar{R}_{cs}$, the proportion of the class c that repeat the match in the subject s. Fixed effects at class level are included to capture all the subject-invariant characteristics (observable and unobservable) of the class.

Results are displayed in Table 2.7. Column (1) indicates that repeat matches have a positive effect on attendance, increasing it by 0.05σ, an effect size which seems plausibly consistent with the effect on test scores. Repeat matches also improve the teacher's perception of classroom behaviour and teacher expectations, shown in columns (2)–(4). In particular, teachers are 4.1 percentage points less likely to have behavioural problems at the beginning of the class and 4.4 percentage points less likely to experience disruptive student behaviour. There are smaller but still significant effects on teacher expectations: teachers are 1.7 percentage points more likely to hold higher expectations for their students if their class is entirely made up of repeated matches.

These results are consistent with the qualitative evidence from teachers who claim that "looping" is beneficial for classroom behaviour. Students are familiar with the expectations of behaviour set by the teacher in previous years, and as a result behaviour improves. Of course, we cannot

[28] 41% of the classes in data have this survey information for each subject. Table B.5 (Appendix B.5) reports a mean comparison test of classroom observable characteristics for the estimation sample and the restricted sample. The restricted sample has more socio-economically advantaged students, and also has students with a better average performance in the SIMCE test. Although the differences between the two samples are statistically significant, they are not large.

Table 2.7. Effect of repeat matches on student behaviour and teacher expectations

	Attendance	Problems to start the class	Classroom disruption	High teacher expectations
	(1)	(2)	(3)	(4)
Proportion of repeat matches	0.052***	-0.041***	-0.044***	0.017**
	(0.008)	(0.007)	(0.008)	(0.007)
Class FE	Yes	Yes	Yes	Yes
Subject FE		Yes	Yes	Yes
Student controls	Yes			
Teacher controls		Yes	Yes	Yes
R-squared	0.516	0.418	0.439	0.566
Observations	696,482	37,992	37,992	37,992

Notes: Column (1) is at the student level and includes controls for students' gender, household income, mother's education, and attendance rate in grade 7. Columns (2)–(4) are at the class-subject level and include controls for the teacher's gender and experience. Standard errors are clustered at the class-level. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

tell if the positive effects of repeat matches are jointly responsible for improved student behaviour and improved test scores, or whether improved behaviour is a mechanism by which academic performance improves.

2.6 Conclusions

There is a large literature which stresses the importance of teacher quality for student outcomes. But teacher quality is hard to improve. In this study, we have provided evidence that there are significant benefits to reallocating existing teachers to students they have taught before. Qualitative evidence from teachers suggests that repeating the match saves time, engenders greater familiarity, and hence aids learning. However, estimating the causal effect of student-teacher familiarity is challenging for two reasons. First, because student-teacher matches are non-randomly selected. Second, because, even if student-teacher matches were chosen randomly, a repeat match may affect student performance for reasons other than student-teacher familiarity: we have seen that repeat matches have more experienced teachers and may also have more within-class familiarity.

We have provided a range of evidence from a new setting to suggest that repeating the student-teacher match has a significant positive effect on student test scores: we consider older (grade 8) children in a situation where repeat matches are common. A multidimensional fixed-effects framework which controls for selection by student or teacher into repeat matches suggests that repeat matches have test scores about 0.02σ higher, a result which is very consistent with evidence for younger children from the US. Our results also support a wide range of case-study and qualitative findings from the educational literature. The fixed-effects methods effectively hold constant many of the other channels by which repeat matches might affect student outcomes. A regression discontinuity design which additionally controls for selection on the basis of subject-specific match quality suggests larger effects in the range 0.11σ to 0.16σ, albeit with much less precision.

We have also shown that these effects aggregate to the class and school-level, which implies that the positive effects for treated classes are not simply at the expense of untreated classes, which would be the case if, for example, schools simply allocate more effective teachers to repeat matches. Consistent with our findings of positive effects on test scores, we also find positive effects in teachers' perceptions of classroom behaviour and their expectations of their students' achievements.[29]

Allocating teachers to groups of students with whom they have interacted in the past appears to bring significant improvements in student performance without incurring additional costs on schools. An important question for future research is whether these results, which are estimated from variation in repeat matches in observational data, can be verified in a randomised setting.

[29] Note that our measure of test scores comes from an anonymous national test which is not marked by the teacher, so there is no mechanistic relationship between test scores and teachers' perceptions.

Chapter 3

School starting age and sibling spillover effects on college admission exams

3.1 Introduction

A substantial research literature has established the role of parents (Todd and Wolpin, 2007; Cunha and Heckman, 2008), teachers (Rivkin et al., 2005; Chetty et al., 2014*a*; Chetty et al., 2014*b*) and peers (Sacerdote, 2011; Epple and Romano, 2011) on the formation of human capital. So far, however, there has been little discussion about the impact of siblings on student academic outcomes. Siblings naturally share genes, parental characteristics, parental resources, family values, family history, and neighbourhoods; and simultaneously compete for parental attention and investments. They are also likely to influence each other's development when they act as social partners and role models, and as the focus of social comparisons (Dunn, 2007; Whiteman et al., 2011; McHale et al., 2012). However, it has proved difficult to establish the causal effect due to the reflection problem and unobserved correlated factors (Manski, 1993; Blume et al., 2011). Because random assignment of siblings is not possible, researchers have used quasi-experimental methods to investigate spillovers between siblings.

In this chapter, I combine detailed administrative data from Chile to examine how children affect the long-term educational outcomes of their siblings.

In particular, this study asks whether children's school starting age has spillover effects on their siblings. To identify these effects, I exploit an exogenous variation in children's school starting age caused by Chile's school entry cutoff date. I link data on college admission test scores between 2004-2019 to students' exact birth date and enrolment age. I complement this dataset with information on school achievement (SIMCE test scores) and survey data on parental investments reported by children and parents.

I obtain three main results. First, school starting age has significant effects on in-school outcomes, college entrance exams and college enrolment. In particular, children who enter school at an older age have better academic outcomes than their classmates. These results match those observed in earlier studies. Furthermore, I find that these effects are smaller for older siblings. Second, I provide evidence of spillovers from younger-to-older siblings. Children score 0.05σ higher in college admission exams if their younger sibling enters school at an older age. In other words, the advantage that children gain from starting school older has a positive effect on their older siblings. Surprisingly, this advantage does not spillover to their younger siblings. To put this result in context, the effect is almost a third of the gender gap in math test scores and almost a fifth of the gap in the average of mathematics and language test scores between public and voucher schools. Similar results are obtained using SIMCE test scores in 8th (students aged 13–14) and 10th grade (students aged 15–16). Third, I provide direct evidence that parental investments are larger for the older sibling when the younger sibling starts school older, suggesting that parents react to the gains that their younger children receive by allocating more investment to their older children.

This chapter connects to three strands of the literature. First, it is related to the literature on sibling spillovers and the role of the family on educational outcomes. Most of this literature finds spillovers from older to younger siblings (e.g. Dustan, 2018; Joensen and Nielsen, 2018; Dahl et al., 2020; Aguirre and Matta, 2021; Altmejd et al., 2021). For example, Joensen and Nielsen (2018) study sibling spillovers in Denmark exploiting an exogenous variation in the cost of taking advanced courses in math and science in high school. They find that younger siblings are more likely to choose advanced math and science in high school if their older siblings also take these courses. Similarly, Dustan (2018) using score-based admission rules in high school in Mexico City, shows that older siblings' school assignments impact younger siblings' stated preferences as well as their assignment outcomes. Altmejd et al. (2021) using a regression discontinuity design based on college admissions thresholds, present causal evidence from Chile, Croatia, Sweden, and United States that older sibling's college and major choices strongly affect younger sibling's college choice behaviour, increasing their likelihood of applying and enrolling in the same college and major than their older sibling.

The closest works on this topic is by Karbownik and Özek (2019) and Landersø et al. (2020). These two recent studies use discontinuity designs around the minimum school-entry age to investigate intrafamily spillovers and sibling spillovers. Karbownik and Özek (2019) use administrative data on student test scores for 3rd-8th graders from the state of Florida. They compare test scores of children whose sibling was born before and after the school-entry cutoff, finding positive spillovers from older-to-younger siblings; a child who starts school at an older age has a positive effect on their younger sibling. These gains are entirely driven by students from low-SES backgrounds. They also find small negative spillovers from younger-to-older siblings only in the case of students from high-SES backgrounds. They argue that there are two channels through which children's performance could affect the educational outcomes of their siblings. First, there is a direct spillover explained by the interactions between siblings. They expect a positive effect if high-achieving students act as role models, especially for students coming from low-income families. Second, there is an indirect channel through which parents react to differences in their children's performance, either by allocating more resources in the lower performing child (i.e. adopting a compensating strategy) or by allocating more resources in the higher performing child (i.e. adopting a reinforcement strategy). Therefore, these results are consistent with direct spillovers in students coming from lower socioeconomic background and reinforcing behaviour in students coming from higher socioeconomic background. Landersø et al. (2020) using Danish data find that a higher school starting age improves older siblings' test scores at the end of grade 9. They argue that a higher school starting age of younger siblings improves the school performance of the older siblings “because the study environment at home is better or because parental resources are freed to assist with homework”.[1] In support of this argument, they show that only school grades related to rote learning (e.g. basic arithmetic and grammar) – where parents can actively help their children to prepare the tests – improve substantially. In contrast, no effect is found for school grades where more meaningful and deeper learning is needed (e.g. essay writing or the oral examination of text analysis), which is not easily improved in the short run by practice. This study complements and extends their work in at least two important dimensions. First, I focus on the effect of sibling’s school starting age on long-run outcomes, such as college admission exams. Second, I explore behavioural channels that might explain the results. I provide evidence that parents do indeed respond to the timing of child’s school start, as proposed by Landersø et al. (2020).

Second, this chapter also relates to the literature on intra-household resource allocation and its interaction with early life shocks on human capital. Scholars have long debated about how parents allocate resources among their offspring so as to reinforce or compensate initial endowments. In their classic theoretical model of intra-household allocation decisions, Becker and Tomes (1976) argue that parents invest more human capital in better-endowed children, and therefore reinforce endowment differentials among their children. On the other hand, Behrman et al. (1982) posit that parents face an equity-productivity trade off. Then, if equity is weighed more heavily in parental preferences, parents will compensate for endowment differences. The empirical evidence is mixed. To date there has been little agreement on whether parents reinforce or compensate children’s endowment differences. Some studies find evidence that parents reinforce (e.g. Rosenzweig and Zhang, 2009; Datar et al., 2010; Rosales-Rueda, 2014; Frijters et al., 2013), while others find that parents compensate (e.g. Del Bono et al., 2012; Yi et al., 2015; Bharadwaj et al., 2018).[2] These divergent results can be explained by authors using different measures of children’s en-

[1] They do not examine spillovers from older-to-younger siblings.

[2] There are some studies that find that parents are neutral, and do not compensate or reinforce endowment differences (e.g. Royer, 2009; Currie and Almond, 2011).

dowments and/or by differences in the empirical methods. Previous studies have extensively used birth weight as a measure of child endowments (e.g. Royer, 2009; Del Bono et al., 2012; Yi et al., 2015; Restrepo, 2016; Bharadwaj et al., 2018); whereas measures of other domains of endowments (e.g. cognitive skills) have been less used (e.g. Aizer and Cunha, 2012; Frijters et al., 2013; Sanz-de Galdeano and Terskaya, 2019). Moreover, parental responses might change according to the dimensions of human capital. In that regard, Yi et al. (2015) show that parents simultaneously compensate or reinforce in different dimensions of human capital in response to early health shocks. In particular, they find that in response to early health shocks, parents make compensating investments in child health, whereas they make reinforcing investments in education. The estimation of parental responses to children's endowments poses several econometric challenges. The main threat to identification is selection bias caused by unobserved characteristics that might be correlated with both parental investments and children's endowments. To deal with this problem, studies have used different strategies, such as family fixed effects (e.g. Rosales-Rueda, 2014), natural experiments (e.g. Aizer and Cunha, 2012) and instrumental variables (e.g. Yi et al., 2015). Relative to these papers, I contribute by providing evidence consistent with parents acting to compensate the well-being of *worse-endowed* children (i.e. children who enter school at a younger age).[3] For identification, I exploit an arguably exogenous variation caused by school entry laws, where children are randomly assigned to start school based on their date of birth.

Third, this chapter speaks to the extensive literature that examines the relationship between school starting age and academic performance. Previous research has shown that students who enter school at an older age score higher on in-school tests (Bedard and Dhuey, 2006; Crawford et al., 2007; Puhani and Weber, 2007; McEwan and Shapiro, 2008; Nam, 2014; Lubotsky and Kaestner, 2016; Attar and Cohen-Zada, 2018; Dhuey et al., 2019) and are more likely to attend college (Bedard and Dhuey, 2006; Gal-

[3] These results reflect those of Bharadwaj et al. (2018) which also use data from Chile. They examine the relationship between health at birth – measured by birth weight – and school outcomes, finding that birth weight increases outcomes in math and languages throughout all the school years. They show that parents invest more in children with lower birth weight, this suggests that parental investments are compensatory.

legos and Celhay, 2020).[4] It is well known that school starting age may be endogenous and is likely to correlate with student and family characteristics, parental preferences, and child maturity (Attar and Cohen-Zada, 2018; Landersø et al., 2020). To address these concerns some studies exploit discontinuities in school starting age due to birth date cutoff rules, comparing children who born just before the enrolment cutoff with those who born just after (McEwan and Shapiro, 2008; Black et al., 2011; Peña, 2017; Gallegos and Celhay, 2020). I contribute to this literature by showing that older siblings are less affected by the school starting age. On the other hand, and consistent with the literature of birth order effects, I show that first-borns do better on several measures of educational attainment than later-borns (Black et al., 2005; Barclay, 2018).

The remainder of the chapter is organized as follows. Section 3.2 describes the data and how siblings are identified in the sample. Section 3.3 explains the empirical strategy and evaluates the validity of the Regression Discontinuity (RD) design. Section 3.4 presents the main findings of the study. Section 3.5 discusses the potential underlying mechanisms. Finally, section 3.6 concludes.

3.2 Data

I use three different datasets in this study. The first source of data is the enrolment records of the entire student population in Chile between 2004 and 2019. The data come from administrative records provided by the Ministry of Education, and contain information on students' exact date of birth, the year in which they start school and the municipality where they reside. I also have access to students' surnames. Traditionally, Chile follows the Spanish naming system, in which children have two surnames, the

[4] Recent studies have also documented that higher school starting age decreases the incidence of crime in youth (Cook and Kang, 2016; Depew and Eren, 2016; Landersø et al., 2017) and lowers the probability of teenage pregnancy (Black et al., 2011). The literature on job market outcomes is less consistent. Some authors find that students who enter school at an older age have higher wages (Fredriksson and Öckert, 2013; Peña, 2017), while others find no effect of the school starting age on earnings (Dobkin and Ferreira, 2010; Black et al., 2011).

first being the father's first surname followed by the mother's first surname.

The second source of data is from a national standardised test administered by the Ministry of Education to all students in certain grades and years. The test is taken at three different times: 4th grade (9-10 year old students), 8th grade (13-14 year old students) and 10th grade (15-16 year old students). The national exam called *Sistema de Medición de la Calidad de la Educación* (SIMCE) is the main tool to measure the quality of education in Chile. I use standardised test scores spanning 2000–2017 in two subjects: Mathematics and Spanish.[5] The SIMCE test also collects information from surveys of parents and students. The parent survey contains information about the characteristics of the household as well as the parents' perceptions about the school. In 2007, the parent survey for 4th graders included questions about parental investment. Parents were asked about how frequently they engage in certain activities with their children. Specifically, the survey includes the following questions: "How often do you read to your child?", "How often do you read with your child?", "How often do you talk to your child about his reading?" The student survey contains information about parental involvement and parental monitoring. In years 2009, 2011, 2013 and 2014, 8th grade students were asked about how engaged their parents in their school lives are. Students report how often parents congratulate them for their school grades, know their grades in school, and are willing to help them out when they are struggling in school. Parental investments in 4th grade are on a scale of 1-5, where 1 denotes "never" and 5 denotes "very often". These variables are coded as dummy variables, taking value of one if parents report "very often" or "often", and zero otherwise. On the other hand, parental investments in 8th grade are on scale of 1-4, where 1 denotes "I entirely disagree" and 4 denotes "I fully agree". These variables are coded as dummy variables, taking value of one if students respond "I fully agree" or "I agree", and zero otherwise.

[5] I focus on these two subjects because of the availability of SIMCE test score information. The national examination also tested the subjects of social sciences and natural sciences, but these are taken for certain grade-year combinations.

The third source of data is the record of students who register for the *Prueba de Selección Universitaria* (PSU) test after graduating from high school. The PSU test is the college entrance exam and is administered by the *Departamento de Medición, Registro y Evaluación* (DEMRE), which is the agency responsible for admission to higher education in Chile. The test is administered once a year and all the applicants take the test on the same day at the end of the academic year. Applicants take two mandatory tests in Mathematics and Language, and at least one of two optional tests in Social Science and Natural Science. PSU test scores range from 150 to 850, and each test is normalised to have a mean of 500 and a standard deviation of 110. Between 2004 and 2019, 86% of the students who graduated from high school registered for the PSU immediately after graduation, and 91% of them took PSU test. College admission is based on a weighted average of their PSU test scores, high school grades and Grade Point Average (GPA) ranking.[6] The data in this study consider the PSU exams in Mathematics and Language for the students graduated from high school during the years 2004–2019. The PSU dataset also provides information on student characteristics – including gender, monthly family income, parents' schooling, family size, parents' work status and health coverage – and high school GPA. Finally, I also observe college enrolment for those who take the PSU test. This data is available from 2006 onwards.

The three datasets are linked using a unique student identification number. The main analysis focuses on the cohort of students that graduated from high school between 2004 and 2019, and immediately took the PSU test.[7]

3.2.1 Identifying Siblings

I identify siblings using students' surnames provided by the Ministry of Education. I define two students as siblings if they share the same pair of surnames (in the same order), they attend the same school and live in the

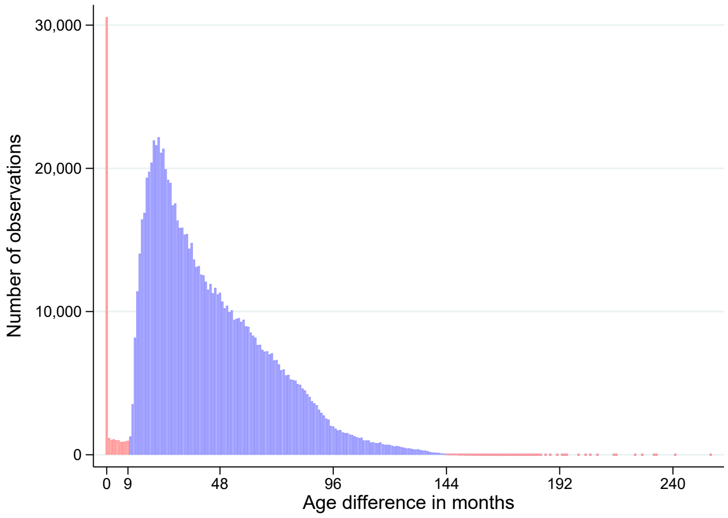
[6] GPA ranking was introduced in 2013 as one of the components of the weighted average for college admissions.

[7] The scores are standardised and have a zero mean and a standard deviation of one. Also, I standardise high school grades at school-level to capture differences in grading standards.

same municipality.[8] Using this strategy, I identify 969,406 pairs of siblings between 2004 and 2019. I drop (1) twins because they have the same date of birth and therefore there is no variation in the treatment (30,568 pairs of siblings), (2) sibling pairs born less than nine months apart (9,097 pairs of siblings), and (3) siblings pairs whose age difference is greater than 12 years because it is not possible to claim that he/she was affected by the shock (421 pairs of siblings).

Figure 3.1 shows the distribution of age differences in months. The existence of pairs with difference in date of birth of less than 9 months (0.9% of total siblings) indicate that might be some pairs of students incorrectly identified as siblings. In any case, this should attenuate the RD estimations towards zero, and work against the existence of sibling spillovers.

FIGURE 3.1. Distribution of age differences (in months)



Notes: The red sections are removed from the sample. Those observations correspond to twins who have the same date of birth, siblings pairs born less than nine months apart, and siblings whose age difference is greater than 12 years.

[8] Aguirre and Matta (2021) implement a similar strategy to study sibling spillovers on higher education choices in Chile. They classify two students as siblings if (i) they share the same pair of surnames (in the same order) and (ii) they go to the same school. To evaluate their method, they match their records with administrative information on parents' national identification numbers – acceded on-site at the Ministry of Education. They conclude that 93% of students in the same school who share the same surnames are siblings.

The estimation sample comprises 260,393 sibling pairs (520,786 students) who have valid PSU test scores and complete information on socioeconomic characteristics. Table 3.1 presents summary statistics. In panel (a) I report information on socioeconomic characteristics of students, including gender, age gap between siblings, whether their sibling has the same gender, whether lives in the capital, monthly family income, parents' schooling, family size, parents' work status and health coverage. Panel (b) shows information on academic performance: high school GPA, PSU test scores (Mathematics and Language) and college enrolment, and SIMCE test scores in 4th grade, 8th grade, and 10th grade.[9] In panel (c) I report information on parental investments in 4th grade using responses from surveys of parents. Finally, in Panel (d) I report information on parental investments in 8th grade using responses from surveys of students.

Table 3.2 shows same-age comparisons between younger and older siblings. In Panel (a) I show the differences in socioeconomic characteristics. By definition, siblings share some characteristics, and therefore present no differences in age gap between siblings, whether their sibling has the same gender, whether lives in the capital, monthly family income, parents' schooling, and family size. By contrast, parents' work status and health coverage might vary between siblings because it reflects the conditions at the end of grade 12 when students must register for the PSU test. The father is less likely to be working at the end of grade 12 for older siblings vis-à-vis younger siblings, while the opposite is true for mother's employment status. On the other hand, younger siblings are more likely to be covered by a public health plan. Panel (b) shows that older siblings exhibit a better academic performance than younger siblings, in terms of school grades, PSU test scores, college enrolment and SIMCE test scores. Panel (c) and (d) show that older siblings tend to receive more attention and help with their school work from their parents than younger siblings.

[9] The information on SIMCE test scores is available for a sub-sample of students (see Appendix C.1). Also, I do not consider SIMCE test scores for firstborns whose siblings are not yet in school, because they are not affected by their younger sibling's school starting age.

Table 3.1. Descriptive statistics

	Mean	Standard deviation	Observations
(a) Student characteristics			
1=Female	0.52	0.50	520,786
Age difference	3.58	1.99	520,786
1=Same-sex sibling	0.54	0.50	520,786
1=Lives in the capital	0.42	0.49	520,786
Monthly family income	563.05	509.50	520,786
Mother's schooling	12.44	3.37	520,786
Father's schooling	12.51	3.61	520,786
Family size	4.53	1.13	520,786
1=Father employed	0.83	0.37	393,455
1=Mother employed	0.46	0.50	393,455
1=Public health coverage	0.65	0.48	393,455
(b) Academic performance			
High school GPA	56.91	5.01	520,786
PSU: Math score	517.47	111.02	520,786
PSU: Language score	511.34	108.97	520,786
1=College enrolment	0.45	0.50	488,671
SIMCE 4th grade	0.46	0.92	215,322
SIMCE 8th grade	0.47	0.88	224,912
SIMCE 10th grade	0.56	1.02	288,309
(c) Parental investments: 4th grade			
1=Read to child	0.35	0.48	19,835
1=Parent-child joint reading	0.52	0.50	19,835
1=Talk about their readings	0.61	0.49	19,835
(d) Parental investments: 8th grade			
1=Parent congrats for grades	0.85	0.35	99,579
1=Parent knows grades in school	0.81	0.39	99,579
1=Parent willing to help	0.82	0.39	99,579

Notes: The estimation sample comprises students who have valid PSU test scores between the years 2014 and 2019, and a complete set of information on characteristics. The information on whether parents are employed (*Father employed* and *Mother employed*) and health coverage status (*Public health coverage*) is only available for 393,794 students (76% of the sample). Enrolment information is only available from 2006 onwards. Information on SIMCE test scores – in panel (b) – and parental investments – in panels (c) and (d) – is only available for a sub-sample of students (see Appendix C.1).

TABLE 3.2. Mean comparison test: younger and older siblings

	Younger siblings		Older siblings		Difference	Std. err.
	Mean	Obs.	Mean	Obs.		
(a) Student characteristics						
1=Female	0.52	260,393	0.52	260,393	0.002	(0.001)
Age difference	3.58	260,393	3.58	260,393	0.000	(0.006)
1=Same-sex sibling	0.54	260,393	0.54	260,393	0.000	(0.001)
1=Lives in the capital	0.42	260,393	0.42	260,393	0.000	(0.001)
Monthly family income	563.05	260,393	563.05	260,393	0.000	(1.412)
Mother's schooling	12.44	260,393	12.44	260,393	0.000	(0.009)
Father's schooling	12.51	260,393	12.51	260,393	0.000	(0.010)
Family size	4.53	260,393	4.53	260,393	0.000	(0.003)
1=Father employed	0.82	166,174	0.85	227,281	-0.031***	(0.001)
1=Mother employed	0.47	166,174	0.45	227,281	0.017***	(0.002)
1=Public health coverage	0.65	166,174	0.64	227,281	0.014***	(0.002)
(b) Academic performance						
High school GPA	56.85	260,393	56.96	260,393	-0.105***	(0.014)
PSU: Math score	515.94	260,393	519.01	260,393	-3.071***	(0.308)
PSU: Language score	508.16	260,393	514.52	260,393	-6.357***	(0.302)
1=College enrolment	0.44	258,274	0.47	230,397	-0.026***	(0.001)
SIMCE 4th grade	0.45	161,177	0.51	54,145	-0.062***	(0.005)
SIMCE 8th grade	0.43	129,958	0.53	94,954	-0.100***	(0.004)
SIMCE 10th grade	0.54	161,859	0.59	126,450	-0.051***	(0.004)
(c) Parental investments: 4th grade						
1=Read to child	0.33	14,026	0.39	5,809	-0.065***	(0.007)
1=Parent-child joint reading	0.51	14,026	0.53	5,809	-0.023***	(0.008)
1=Talk about their readings	0.60	14,026	0.62	5,809	-0.022***	(0.008)
(d) Parental investments: 8th grade						
1=Parent congrats for grades	0.85	71,769	0.86	27,810	-0.008***	(0.003)
1=Parent knows grades in school	0.81	71,769	0.82	27,810	-0.002	(0.003)
1=Parent willing to help	0.82	71,769	0.82	27,810	0.002	(0.003)

Notes: Table shows same-age comparisons between younger and older siblings. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

3.3 Empirical strategy

The estimation of peer effects is challenging due to the well-known problems of (i) endogenous peer group formation, (ii) the reflection bias (iii) and the existence of unobserved correlated effects (also known as common shocks problem) (Manski, 1993; Blume et al., 2011). The problem of endogenous peer group formation arises when individuals self-select into a group, because of the characteristics or the choices of this group. In the case of sibling peer effects this should not be important, because siblings are born in the same family (peer group) and do not choose to be part of that family based on each other's characteristics. The reflection bias states that peer effects are by nature simultaneous, i.e. individuals are affected by peers and vice versa. Finally, the estimation of peer effects is also affected by the presence of unobserved correlated effects. This arises when individuals are similar in terms of unobserved characteristics and these characteristics affect or correlate with the outcomes.

To address these issues, I exploit a discontinuity generated by Chile's school enrolment eligibility criteria to estimate the causal effects of school starting age. A child is eligible to enter school in year t if she turns six years old by July 1 of year t. I compare outcomes of students born before and after the cutoff date, using a RD design. This approach will produce a causal estimate as long as the dates of birth are random near the eligibility threshold. This strategy addresses the reflection problem and the common shocks problem since the variation in the school starting age comes only from being born before and after the cutoff date.

Let Y_i denote the outcome of student i; B_i denotes the date of birth of student i (normalised relative to the eligibility threshold such that students with positive B starts the school at an older age); $\mathbf{1}(B_i \geq 0)$ is an indicator function for whether student i born on or after the cutoff date; $f(B_i)$ is a continuous function of date of birth of student i; and ξ_i is a mean zero error. I model the relationship between starting school at an older age and the outcomes of interest using the following reduced-form equation:

$$Y_i = \beta_0 + \beta_1 \times \mathbf{1}(B_i \geq 0) + f(B_i) + \mathbf{1}(B_i \geq 0) \times f(B_i) + \xi_i \qquad (3.1)$$

The parameter β_0 captures the expected value of Y_i for students whose birthday is just before the cutoff date. The parameter β_1 corresponds to the effect of delaying enrolment and starting school at an older age. I implement a standard RD approach to estimate β_1.

I focus on estimating *intention-to-treat* (ITT) effects using reduced-form regressions. Since date of birth (B_i) is arguably randomly assigned (formally, $\{Y_i, Y_i(0), Y_i(1)\} \perp\!\!\!\perp \mathbf{1}(B_i \geq 0)$), the ITT effects capture the causal effect of the offer of treatment by comparing the average Y_i among students assigned to treatment and students assigned to control.

Then, following the causal inference literature (Rubin, 1974; Holland, 1986), the observed outcome for student i is:

$$Y_i = (1 - T_i) \times Y_i(0) + T_i \times Y_i(1) \tag{3.2}$$

where $T_i \equiv \mathbf{1}(B_i \geq 0)$ denotes the assignment to the treatment; $Y_i(0)$ and $Y_i(1)$ correspond to the potential outcomes that would be observed if student i had been born before and after the cutoff date, respectively. As it is shown in Hahn et al. (2001), if $\mathbf{E}\left[Y_i(1)|B_i = 0\right]$ and $\mathbf{E}\left[Y_i(0)|B_i = 0\right]$ are continuous functions of B_i at $B_i = 0$, the treatment effect β_1 can be identified by:

$$\beta_1 \equiv \mathbf{E}\left[Y_i(1) - Y_i(0)|B_i = 0\right] = \lim_{B \downarrow 0} \mathbf{E}\left[Y_i|B_i = 0\right] - \lim_{B \uparrow 0} \mathbf{E}\left[Y_i|B_i = 0\right] \tag{3.3}$$

I adopt a local-linear approach to approximate the unknown functions in Equation 3.1, using a triangular kernel function to weight the observation between a bandwidth $h \in [-h, +h]$. To select the bandwidth, I follow the procedure proposed by Calonico et al. (2014) by selecting the parameter h that minimises an approximation to the asymptotic mean squared error of the point estimator. Finally, for inference I use robust standard errors and confidence intervals proposed by Calonico et al. (2014).

To study the spillover effects of siblings' school starting age, I compare outcomes of students whose siblings born right before and right after the school-entry cutoff. I estimate a modified version of Equation 3.1, where the threshold now refers to the threshold of child i's sibling:

$$Y_i = \gamma_0 + \gamma_1 \times \mathbf{1}(B_j \geq 0) + f(B_j) + \mathbf{1}(B_j \geq 0) \times f(B_j) + \nu_i \tag{3.4}$$

where Y_i is the outcome of student i; B_j denotes the date of birth of student $j = J(i)$, where $J(\cdot)$ is a function that maps student i to their sibling j; $\mathbf{1}(B_j \geq 0)$ is an indicator function for whether student j born on or after the cutoff date; $f(B_j)$ is a continuous function of date of birth of student j; and ν_i is a mean zero error. The parameter of interest is γ_1, which represents the causal effect of a sibling's enrolment cutoff on a child outcome. I use the method presented above to estimate γ_1. Equation 3.4 allows to investigate the spillover effects from older to younger siblings, as well as the spillover effects from younger to older siblings.[10]

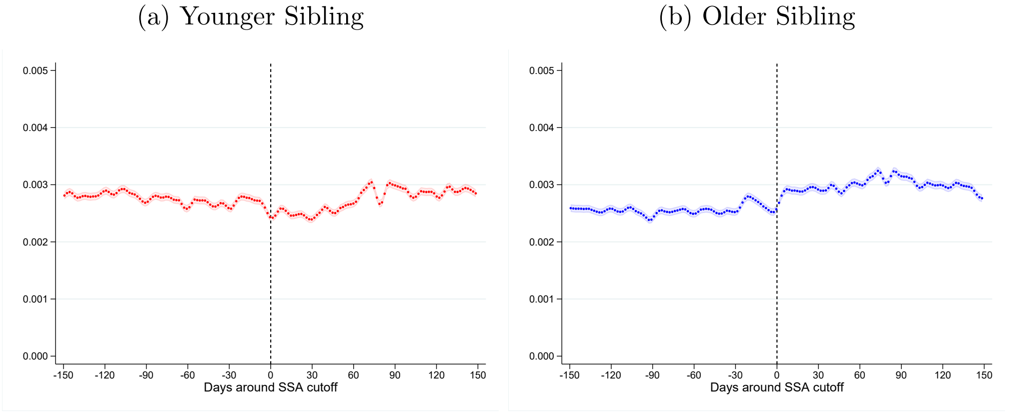
3.3.1 Validity of the discontinuity

A RD design will produce unbiased estimates of the treatment only if there is no manipulation of the running variable around the threshold. To empirically test the validity of the RD design, I consider two falsification tests. First, I examine whether parents self-select into treatment by manipulating their date of birth, causing a jump in the density after the cutoff date. Figure 3.2 shows a kernel density estimate of the date of birth for younger (Panel 3.2a) and older siblings (Panel 3.2b). The graphical evidence suggests that the empirical density is continuous across the threshold. To confirm this, I implement a formal test suggested by Cattaneo et al. (2018), with the null hypothesis that the there is no jump in the density on July 1. In the case of younger siblings the p-value is 0.40, whereas in the case of older siblings the p-value is 0.19. Therefore, we cannot reject the null hypothesis of no differences in the density of treated and control observations at the eligibility cutoff. Overall, these results indicate that there is no evidence of manipulation of the running variable at the cutoff.

Second, I test whether the observable characteristics of students are balanced on the two sides of the school-entry cutoff. If there is a non-random sorting around assignment cutoff, it should be expected that students on one side of the threshold to be systematically different from those on the other side. In order to implement the analysis, I estimate reduced-form regressions using each covariate as the outcome variable. I use information available on student characteristics, which includes gender, whether lives in

[10] A potential concern is that the date of birth of the student (B_i) might be correlated with the date of birth of their sibling (B_j), in which case B_i (B_j) is an omitted variable in Equation 3.4 (Equation 3.1). However, a regression of the date of birth of the student i (B_i) on the date of birth of their sibling (B_j) – controlling for student characteristics (gender, whether lives in the capital, monthly family income, mother's schooling, father's schooling, and family size) – shows that there is no correlation between the dates of birth. The regression yields an estimate of 0.002 (0.002).

Figure 3.2. Estimated density of date of birth relative to school-entry cutoff date



Notes: Panels (a) an (b) show an estimated density of date of birth within -/+150 days with shaded 95% confidence intervals using 200 points and a bandwidth of 5 days, for younger siblings and the older siblings, respectively. The dashed line corresponds to 1st of July.

the capital, monthly family income, mother's schooling, father's schooling, family size, whether parents are employed and health coverage status.[11] I report the regression results in Panel (a) of Table 3.3. There are no discontinuities in pre-determined variables: there are no differences between the characteristics of students born before and after the school-entry cutoff. The balance of pre-determined variables on either side of the cutoff holds when the analysis is performed separately for younger (Panel (b) of Table 3.3) and older siblings (Panel (c) of Table 3.3). Furthermore, to test the presence of sibling spillovers, I investigate whether covariates are also "locally" balanced on the two sides of the threshold using sibling's school starting age cutoff as a running variable. Table 3.4 presents the balancing tests performed separately for younger (Panel (a) of Table 3.4) and older siblings (Panel (b) of Table 3.4). The results indicate that student characteristics are balanced on both sides of the threshold.

[11] The information on whether parents are employed (*Father employed* and *Mother employed*) and health coverage status (*Public health coverage*) is only available for 393,794 students (76% of the sample).

Table 3.3. Balance on pre-determined covariates

Covariate	RD estimator (1)	Std. err. (2)	Obs. left (3)	Obs. right (4)	Bandwidth (5)	Observations (6)
(a) All students						
1=Female	-0.003	0.006	85,218	89,729	63.26	520,786
1=Lives in the capital	0.000	0.006	50,529	52,485	37.99	520,786
Monthly family income	7.649	4.985	55,642	58,256	41.17	520,786
Mother's schooling	0.008	0.042	58,353	61,153	43.28	520,786
Father's schooling	0.018	0.061	59,725	62,645	44.64	520,786
Family size	0.004	0.020	67,879	70,828	50.60	520,786
1=Father employed	-0.004	0.006	70,425	75,184	69.34	393,455
1=Mother employed	0.005	0.009	48,061	50,372	47.24	393,455
1=Public health coverage	-0.000	0.008	60,326	63,748	59.18	393,455
(b) i: Younger sibling						
1=Female	-0.009	0.008	41,984	40,119	61.93	260,393
1=Lives in the capital	0.007	0.007	38,718	36,815	56.47	260,393
Monthly family income	10.319	8.009	40,117	38,178	58.56	260,393
Mother's schooling	0.043	0.064	23,504	22,240	34.87	260,393
Father's schooling	0.039	0.074	28,900	27,585	42.87	260,393
Family size	-0.009	0.023	33,143	31,417	48.25	260,393
1=Father employed	-0.002	0.008	27,235	25,880	62.72	166,174
1=Mother employed	-0.001	0.012	30,936	29,803	70.91	166,174
1=Public health coverage	-0.007	0.011	29,082	27,762	66.35	166,174
(c) i: Older sibling						
1=Female	0.001	0.009	40,508	46,513	61.51	260,393
1=Lives in the capital	-0.005	0.010	32,065	36,537	48.52	260,393
Monthly family income	4.918	8.588	31,422	35,776	47.14	260,393
Mother's schooling	-0.058	0.051	38,662	44,299	58.21	260,393
Father's schooling	-0.001	0.075	33,394	38,069	50.85	260,393
Family size	0.015	0.023	33,394	38,069	50.02	260,393
1=Father employed	-0.008	0.008	40,521	46,643	70.73	227,281
1=Mother employed	0.007	0.011	23,457	26,455	40.53	227,281
1=Public health coverage	0.009	0.009	26,774	30,296	46.06	227,281

Notes: Table shows the estimated discontinuity in each covariate at the threshold. The running variable corresponds to the student's date of birth relative to the eligibility threshold. Results based on the empirical strategy that implements a RD following the methodology proposed by Calonico et al. (2014), with a polynomial of order one and weighted by triangular kernel. Standard errors are clustered at the running variable at daily level. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

3.4 Results

3.4.1 Discontinuities in enrolment age

The Ministry of Education establishes the school entry cutoff date on the 1st of July, but schools are free to implement any cutoff before that date. However, the latest enrolment cutoff – i.e. 1st of July – is the most common enrolment rule.[12] Figure 3.5 plots the enrolment age in grade 1 against

[12] I analyse how prevalent among the students and the schools is the 1st of July as an enrolment cutoff, using data on 1st grade students between 2004-2017. The sample comprises a population of 3,335,764 students enrolled in 9,685 schools. The compliance with the July 1 cutoff is almost universal: 99.9% of the students enrolled in grade 1 turn six years old before July 1. At the school level, 76.7% of the schools systematically use July 1 as their enrolment cutoff; followed by June 1 (8.7%), March 1 (5.3%) and May 1 (3.8%). 529 schools (5.5%) do not comply with the July 1 rule at some point.

Table 3.4. Balance on pre-determined covariates (sibling spillovers)

Covariate	RD estimator (1)	Std. err. (2)	Obs. left (3)	Obs. right (4)	Bandwidth (5)	Observations (6)
(a) Older-to-younger						
1=Female	0.005	0.009	50,414	58,752	76.16	260,393
1=Lives in the capital	-0.005	0.010	32,065	36,537	48.52	260,393
Monthly family income	4.918	8.588	31,422	35,776	47.14	260,393
Mother's schooling	-0.058	0.051	38,662	44,299	58.21	260,393
Father's schooling	-0.001	0.075	33,394	38,069	50.85	260,393
Family size	0.015	0.023	33,394	38,069	50.02	260,393
1=Father employed	0.005	0.008	24,005	28,023	57.74	166,174
1=Mother employed	0.006	0.010	17,347	20,268	41.41	166,174
1=Public health coverage	0.013	0.012	22,776	26,495	54.69	166,174
(b) Younger-to-older						
1=Female	0.003	0.008	22,135	20,937	32.45	260,393
1=Lives in the capital	0.007	0.007	38,718	36,815	56.47	260,393
Monthly family income	10.319	8.009	40,117	38,178	58.56	260,393
Mother's schooling	0.043	0.064	23,504	22,240	34.87	260,393
Father's schooling	0.039	0.074	28,900	27,585	42.87	260,393
Family size	-0.009	0.023	33,143	31,417	48.25	260,393
1=Father employed	0.001	0.006	31,765	30,483	53.41	227,281
1=Mother employed	0.008	0.012	25,067	24,224	42.35	227,281
1=Public health coverage	-0.007	0.010	37,566	36,524	63.16	227,281

Notes: Table shows the estimated discontinuity in each covariate at the threshold. The running variable corresponds to the sibling's date of birth relative to the eligibility threshold. Results based on the empirical strategy that implements a RD following the methodology proposed by Calonico et al. (2014), with a polynomial of order one and weighted by triangular kernel. Standard errors are clustered at the running variable at daily level. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

date of birth. Each dot is the average enrolment age for all students within each day of birth. There are clear increases in enrolment age at $\overline{\mathbf{B}} = \{-90, -60, -30, 0\}$ – which corresponds to the first days of April, May, June, and July, respectively. This figure suggests discontinuities of about 0.5 years at $\overline{B} = \{0\}$, and small increases at $\overline{B} = \{-90\}$, $\overline{B} = \{-60\}$ and $\overline{B} = \{-30\}$. To formally investigate the presence of discontinuities in enrolment age, I estimate the following regression model:

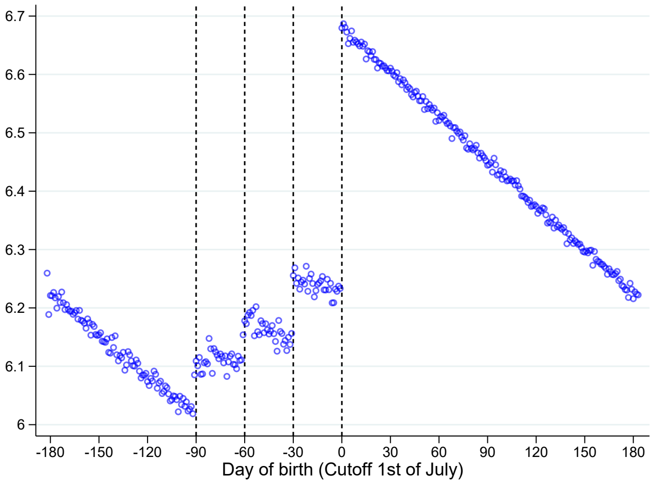
$$\text{SSA}_i = \gamma_0 + \sum_{k=1}^{4} \gamma_k \times \mathbf{1}(B_i \geq \overline{B}_k) + g(B_i) + X_i + \nu_i \tag{3.5}$$

where SSA_i is the school starting age of student i; $\mathbf{1}(B_i \geq \overline{B}_k)$ is an indicator function for whether student i born on or after the cutoff date $\overline{B}_k$, with $\overline{B}_1 = \{90\}$, $\overline{B}_2 = \{-60\}$, $\overline{B}_3 = \{-30\}$, $\overline{B}_4 = \{0\}$; $g(B_i)$ is a piecewise linear polynomial[13]; X_i is a covariate set that includes student characteristics (gender, whether lives in the capital, monthly family income, mother's schooling, father's schooling and family size) and birth-year dummies; and

[13] Defined as: $g(B_i) \equiv \theta_0 \times B_i + \sum_{k=1}^{4} \theta_k \times \mathbf{1}(B_i \geq \overline{B}_k) \times (B_i - \overline{B}_k)$.

ν_i is a mean zero error. Table 3.5 presents the estimates of Equation 3.5. The results confirm the discontinuities at $\overline{\mathbf{B}} = \{-90, -60, -30, 0\}$. Students born in June – just before the eligibility threshold at $\overline{B} = \{0\}$ – start school 0.45 years younger than students born in July.

Figure 3.3. Enrolment age and date of birth



Notes: The dots are mean values of the school starting age within each day of birth.

3.4.2 The effect of own school starting age on PSU test scores

This section presents evidence on the effects of school starting age on college admission exams. Figure 3.4 shows how PSU test scores change with the school-entry cutoff for younger (Panel (a)) and older siblings (Panel (b)). Consistent with the literature, children born after the 1st of July tend to score higher on standardised exams, even when it comes to tests taken 12 years after school enrolment. In addition, the effect of the school entry cutoff date is larger for younger siblings vis-à-vis older siblings.

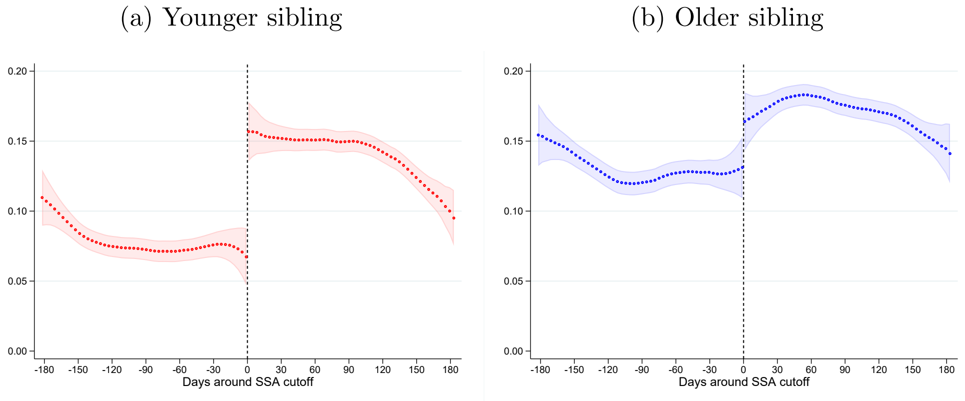
Table 3.6 reports the regression results. Column (1) reports reduced-form effects using a non-parametric model following the methodology proposed by Calonico et al. (2014). The results confirm the visual analysis. The results are quite revealing in several ways. First, they suggest that delay-

TABLE 3.5. Discontinuities in enrolment age at different birth date cutoffs

	(1)	(2)	(3)	(4)
$\mathbf{1}(B_i \geq -90)$	0.090***	0.088***	0.079***	0.097***
	(0.006)	(0.006)	(0.006)	(0.011)
$\mathbf{1}(B_i \geq -60)$	0.070***	0.070***	0.072***	0.066***
	(0.007)	(0.007)	(0.008)	(0.009)
$\mathbf{1}(B_i \geq -30)$	0.115***	0.113***	0.114***	0.112***
	(0.006)	(0.006)	(0.007)	(0.009)
$\mathbf{1}(B_i \geq 0)$	0.454***	0.444***	0.476***	0.411***
	(0.004)	(0.004)	(0.005)	(0.006)
Birth-year dummies		Yes	Yes	Yes
Controls		Yes	Yes	Yes
R-squared	0.213	0.299	0.329	0.271
Observations	520,786	520,786	260,393	260,393

Notes: Columns (2)–(4) control for student characteristics (gender, whether lives in the capital, monthly family income, mother's schooling, father's schooling, and family size) and birth-year dummies. Columns (3) and (4) present the estimates for younger and older siblings, respectively. Standard errors are clustered at the running variable at daily level. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

FIGURE 3.4. Conditional mean plots by local linear regressions: PSU test scores at the threshold


(a) Younger sibling

(b) Older sibling

Notes: Panels (a) an (b) show the discontinuity of the average PSU test score at the threshold for younger and older siblings, respectively. The dashed line corresponds to 1st of July. The graphs show conditional mean plots using local linear regression with shaded 95% confidence intervals within a MSE-optimal bandwidth (Panel (a) bandwidth: 52 days; Panel (b) bandwidth: 49 days), with triangle kernel function and a 1st order polynomial, on a grid of 50 points on each side of the cutoff.

ing school enrolment increases PSU test scores. Column (2) presents RD estimates controlling for student characteristics (gender, whether lives in the capital, monthly family income, father's schooling, mother's schooling, and household size). The inclusion of covariates does not significantly affect the RD estimates. The stability of the estimates can be interpreted as evidence that the no-manipulation assumption holds (Lee and Lemieux, 2010). Column (3) shows the results of local linear regressions within the optimal bandwidth in Column (2), yielding to similar estimates to the baseline model in Column (1).

However, results from previous section show that there are four clear discontinuities at different enrolment cutoff dates, namely April 1, May 1, June 1, and July 1. Therefore, a model that does not jointly consider all the discontinuities might lead to misspecification bias. To address this further concern, I estimate a model that limits the sample to children born 30 days before and after the July 1 cutoff. In other words, I will be comparing the PSU test scores of students born in June with students born in July. The results are shown in Column (4). Reassuringly, these estimates are close to the RD estimates in Columns (1) to (3).

In addition, the same patterns and discontinuities are observed for other outcomes. Specifically, I examine the effects of age differences on five other outcomes: 4th grade test scores, 8th grade test scores, 10th grade test scores, high school grades and college enrolment (see Appendix C.2).

Finally, the results show that skill gaps at the threshold are larger for younger siblings than for older siblings. In other words, younger siblings are more affected by the school starting age. In particular, using the model reported in Column (4) the effect on PSU test scores is 0.05σ higher for younger siblings relative to older siblings. This difference is statistically significant at the 1 percent level.

Table 3.6. The effect of own school starting age on PSU test scores

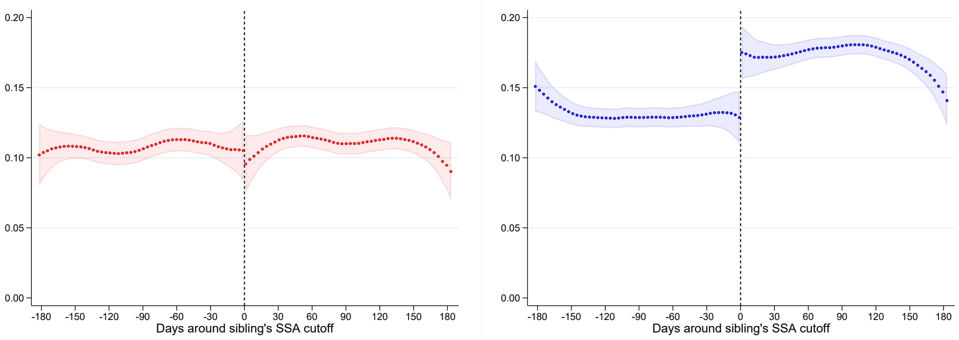
	(1)	(2)	(3)	(4)
Younger sibling:	0.099***	0.099***	0.074***	0.091***
$\mathbf{1}(B_i > 0)$	(0.017)	(0.018)	(0.012)	(0.016)
95% C.I.	[.065 ; .134]	[.064 ; .134]		
Effective obs.: Left	35,190	24,242		
Effective obs.: Right	33,425	22,894		
Optimal Bandwidth	51.98	35.84		
Observations	260,393	260,393	152,632	40,470
Older sibling:	0.031***	0.033***	0.053***	0.039***
$\mathbf{1}(B_i > 0)$	(0.012)	(0.010)	(0.009)	(0.011)
95% C.I.	[.008 ; .053]	[.012 ; .053]		
Effective obs.: Left	32,750	30,088		
Effective obs.: Right	37,263	34,294		
Optimal Bandwidth	49.13	45.08		
Observations	260,393	260,393	155,687	43,437
Controls		Yes	Yes	Yes

Notes: Table shows the effects of school starting age on PSU test scores for younger siblings and older siblings. Results in Columns (1) and (2) are based on the empirical strategy that implements a RD following the methodology proposed by Calonico et al. (2014), with a polynomial of order one and weighted by triangular kernel. Results in Columns (3) and (4) are based on linear regressions within the optimal bandwidth of Column (2) and within +/- 30 days, respectively. Column (1) presents regression with no controls; while Columns (2)-(4) include student characteristics as controls (gender, whether lives in the capital, monthly family income, father's schooling, mother's schooling, and household size). Standard errors are clustered at the running variable at daily level. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

3.4.3 The effect of sibling school starting age on PSU test scores

In this section, I start by investigating the presence of spillovers on college entrance exams generated by birth date cutoff rules. Figure 3.5 plots the PSU test scores of students against their sibling's school starting age cutoff for younger siblings (older-to-younger sibling spillovers) and older siblings (younger-to-older sibling spillovers). This shows that older siblings have higher scores if their sibling was born on or just after the enrolment cutoff. Put differently, older siblings benefit from younger siblings starting school at an older age. By contrast, no effects are found for younger siblings.

Figure 3.5. Conditional mean plots by local linear regressions: PSU test scores at the threshold


(a) Older-to-younger

(b) Younger-to-older

Notes: Panels (a) an (b) show the discontinuity of the average PSU test score at the threshold for younger and older siblings, respectively. The dashed line corresponds to 1st of July. The graphs show conditional mean plots using local linear regression with shaded 95% confidence intervals within a MSE-optimal bandwidth (Panel (a) bandwidth: 48 days; Panel (b) bandwidth: 66 days), with triangle kernel function and a 1st order polynomial, on a grid of 50 points on each side of the cutoff.

The RD estimates presented in Table 3.7 confirm the graphical evidence. The baseline estimate, in Column (1), indicates that older siblings score 0.05σ higher when their siblings born on or just after the cutoff date, whereas no effect is observed for younger siblings. Column (2) presents RD estimates controlling for student characteristics (gender, monthly family income, father's schooling, mother's schooling, and household size). The inclusion of covariates does not significantly affect the RD estimates. Columns (3) and (4) present local-linear regressions within the optimal bandwidth and within +/- 30 days of the 1st of July, respectively. These specifications predict similar gains for older siblings, 0.03σ in Column (3) and 0.06σ in Column (4); whereas suggest no gains for younger siblings. Table C.3 in Appendix C.3 reports the discontinuities in other outcomes: in-school test scores, high school grades and college enrolment. Crossing the enrolment cutoff leads to similar patterns in these outcomes. Except for test scores in 4th grade, older siblings with siblings born on the right-hand side of the eligibility threshold outperform those on the left-hand side. Interestingly, I also find a negative and significant effect on high school grades for younger siblings.

Table 3.7. The effect of sibling school starting age on PSU test scores

	(1)	(2)	(3)	(4)
Older-to-younger:	-0.013	-0.010	-0.005	-0.002
$\mathbf{1}(B_j > 0)$	(0.017)	(0.017)	(0.009)	(0.016)
95% C.I.	[-.047 ; .021]	[-.043 ; .022]		
Effective obs.: Left	32,065	28,785		
Effective obs.: Right	36,537	32,920		
Optimal Bandwidth	48.05	43.43		
Observations	260,393	260,393	153,582	43,437
Younger-to-older:	0.051***	0.052***	0.031***	0.056***
$\mathbf{1}(B_j > 0)$	(0.019)	(0.018)	(0.011)	(0.018)
95% C.I.	[.014 ; .088]	[.017 ; .087]		
Effective obs.: Left	45,516	31,699		
Effective obs.: Right	43,731	30,041		
Optimal Bandwidth	66.26	46.83		
Observations	260,393	260,393	161,155	40,470
Controls		Yes	Yes	Yes

Notes: Table shows the effect of sibling school starting age on PSU test scores for younger siblings (older-to-younger) and older siblings (younger-to-older). Results in Columns (1) and (2) are based on the empirical strategy that implements a RD following the methodology proposed by Calonico et al. (2014), with a polynomial of order one and weighted by triangular kernel. Results in Columns (3) and (4) are based on linear regressions within the optimal bandwidth of Column (2) and within +/- 30 days, respectively. Column (1) presents regression with no controls; while Columns (2)-(4) include student characteristics as controls (gender, whether lives in the capital, monthly family income, father's schooling, mother's schooling, and household size). Standard errors are clustered at the running variable at daily level. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

3.4.4 Heterogeneity in sibling spillovers

In the previous sections, I have shown that sibling school starting age has a positive a significant effect on older siblings' PSU test scores, whereas no spillovers are found in the case of younger siblings. In this section I explore whether these effects differ by student characteristics, in an attempt to uncover underlying mechanisms. The results are presented in Figure 3.6 and Table 3.8. I find no evidence of effect heterogeneity on older-to-younger sibling spillovers (see Columns (1) to (4) in Table 3.8). In contrast, younger-to-older spillovers exhibit substantial heterogeneity (see Columns (5) to (8) in Table 3.8).

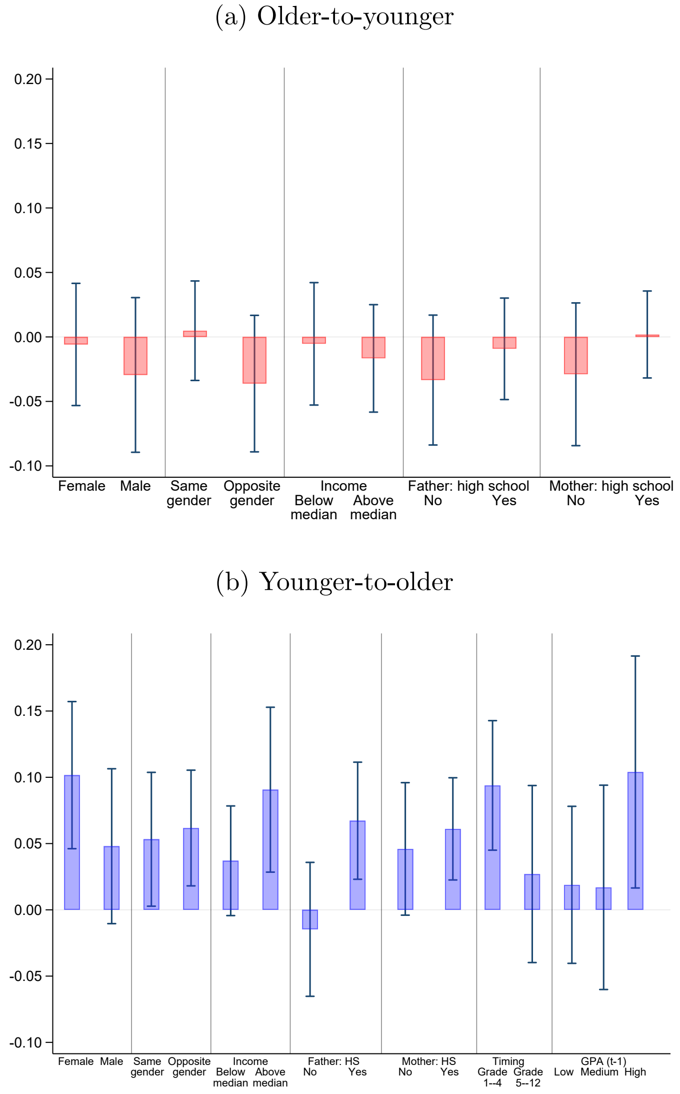
First, I explore differences in younger-to-older spillovers based on the gender of the student. Only female students experience spillovers (0.1σ), while male students are not affected by their sibling school starting age. Then, I test whether sibling spillovers has a stronger impact in same-gender siblings vis-à-vis opposite-gender siblings. I find no systematic differences by siblings' gender composition. Next, I examine differences by student's family background. Students from high-income families with more educated parents tend to exhibit larger spillover effects. Also, I analyse whether sibling spillovers vary with sibling's timing of school start. To test for this, I split the sample into students by grade level: those who were in the first stage of elementary school – between grade 1 and grade 4 – and those who were in higher grades – between grade 5 and grade 12, when their sibling makes the transition into school. The estimates indicate that the impact is statistically significant for those students in grade 1 to 4, and it is insignificant for those students in higher grades. Finally, I test for heterogeneous threshold-crossing effects by school performance one year before their sibling make the transition into school in year t.[14] In particular, I split the sample into three bins of equal size using school GPA in $(t-1)$. Interestingly, the impact is only significant for those students with high levels of performance. I find no sibling spillovers for those students with low or medium performance.

As an additional test of the robustness, I estimate heterogeneous responses using local linear regression within the optimal bandwidth reported in Table 3.8 (see Appendix C.4). Using local linear regressions allows me to test whether the effect significantly varies by student characteristics. This approach delivers remarkably similar estimates. Furthermore, when looking the sibling spillovers running from the younger child to the older, the

[14] The school grades have been standardised at school-grade-year level, to account for school grading standards in a particular grade-year combination. Information about academic performance before their sibling starts school (GPA_{t-1}) is only available for 51% of older siblings. It is worth mentioning that the sibling spillover effect still exists using this sub-sample and is equal to 0.05σ.

results indicate that only the differences by father's schooling and school performance in $(t-1)$ are statistically significant at the 5 percent level, whereas the difference by school timing is statistically significant at the 1 percent level.

Figure 3.6. The effect of sibling school starting age on PSU test scores by student characteristics



(a) Older-to-younger

(b) Younger-to-older

Notes: Panels show heterogeneous effect of sibling school starting age on PSU by student characteristics. Panels (a) and (b) shows these effects for younger siblings (older-to-younger) and older siblings (younger-to-older), respectively. These figures are based on the empirical strategy proposed by Calonico et al. (2014), with a polynomial of order one and weighted by triangular kernel.

Table 3.8. The effect of sibling school starting age on PSU test scores by student characteristics

	Older-to-younger				Younger-to-older			
	RD (1)	Std. err. (2)	Bandwidth (3)	Obs. (4)	RD (5)	Std. err. (6)	Bandwidth (7)	Obs. (8)
Gender								
Female	-0.001	0.022	67.81	136,260	0.097^{***}	0.029	36.63	135,793
Male	-0.027	0.031	57.98	124,133	0.047	0.031	48.43	124,600
Gender composition								
Same-gender	-0.031	0.027	48.79	140,554	0.053^{**}	0.021	60.35	140,554
Opposite-gender	0.006	0.020	51.99	119,839	0.054^{**}	0.026	65.10	119,839
Income								
Below median	-0.005	0.025	45.04	129,642	0.039^{*}	0.021	46.12	129,642
Above median	-0.020	0.022	49.54	130,751	0.080^{***}	0.031	53.82	130,751
Father completed HS								
No	-0.036	0.027	45.42	73,759	-0.015	0.026	66.92	73,759
Yes	-0.008	0.020	50.79	186,634	0.061^{***}	0.023	69.81	186,634
Mother completed HS								
No	-0.028	0.029	55.66	70,746	0.045^{*}	0.026	58.87	70,746
Yes	0.003	0.018	57.23	189,647	0.054^{***}	0.020	65.99	189,647
Timing								
Grade 1 – 4					0.087^{***}	0.028	34.86	156,713
Grade 5 – 12					0.027	0.035	40.83	103,680
GPA$_{t-1}$								
Low					0.019	0.030	40.24	44,200
Medium					0.016	0.039	64.25	44,201
High					0.101^{**}	0.045	43.67	44,199

Notes: Table shows show the heterogeneous effects for younger (older-to-younger) and older siblings (younger-to-older) by student characteristics. Results are based on the empirical strategy that implements a RD following the methodology proposed by Calonico et al. (2014), with a polynomial of order one and weighted by triangular kernel. Standard errors are clustered at the running variable at daily level. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

3.5 Potential channels

The results reveal a significant spillover effect from younger-to-older siblings. Older siblings do significantly better in college admission exams if their sibling starts school at an older age and delays their school enrolment. In contrast, there are no spillovers in the opposite direction, i.e. from older-to-younger siblings. In this section, I explore the potential mechanisms underlying these effects.

One mechanism that may be important is the timing of the enrolment "shock". By definition, for older siblings the shock happens when younger siblings are not yet at school, but for younger siblings it takes place while the older sibling is already at school. This could lead to different parental responses according to which child is affected. For example, if the older child struggles at school in grade 1 – because she starts school younger than her classmates – it does not immediately have any effect on the younger child because she has not yet started school. However, when the younger child struggles at school, it indirectly affects how much time parents can devote to the older child, because parents have limited time to help their children with school work. Table 3.9 shows the relationship between parental investments – reported by parents in grade 4 and by students in grade 8 – and sibling school starting age.[15] These results confirm the timing hypothesis. Column (1) reports the effects on parental investments for younger siblings (i.e. spillovers from older-to-younger). The RD estimates suggest no cross-threshold changes on parental investments. Column (2) reports the effect on parental investments for older siblings (i.e. spillovers from younger-to-older). The results reveal a positive relationship between parental investments and sibling school starting age. In other words, parents invest more in the older child when the younger delayed school enrolment.

It is also possible that parental investment could in part explain the heterogeneous effects found in the previous section. First, we might expect that parental investments are contingent on the age of the student. For example, if the older child is in a higher grade when their sibling starts school, it can be expected that the "loss" in terms of help or motivation from their parents is limited, compared to the case in which the older child is in a lower grade and needs more help in her school work. Second, we would expect that changes in parental investments, because of younger sibling's transition into primary school, are not independent from the performance of the older child. In particular, if before the "shock" the older child is

[15] Information about parental investment in grade 4 and grade 8 is only available for a limited sample. As a robustness check, I show the RD estimates of the main outcome (PSU test scores) using these two samples. Because of the sample size the results are less precise. Nevertheless, the results confirm positive and significant spillovers from younger-to-older siblings, and null spillovers from older-to-younger siblings. On the other hand, in Appendix C.5 I show the results of Table 3.9 using the raw measures of parental investment variable. Investments are on a scale of 1 to 5 for responses from parents in fourth grade, and on a scale of 1 to 4 for responses from students in eighth grade. The results in grade 8 are robust to the way the students' answers are grouped: older siblings tend to receive more attention from their parents when their sibling starts the school older. I find positive – but imprecise – effects using parental responses in 4th grade.

Table 3.9. Parental investments and sibling school starting age

	Older-to-younger	Younger-to-older
	(1)	(2)
(a) 4th grade (Parent Responses)		
1=Read to child	-0.019	0.060
	(0.037)	(0.063)
1=Parent–child joint reading	-0.057	0.124*
	(0.043)	(0.064)
1=Talk about their readings	-0.006	0.134***
	(0.037)	(0.047)
PSU test scores	-0.075	0.191*
	(0.077)	(0.107)
Observations	14,026	5,809
(b) 8th grade (Student Responses)		
1=Parent congrats for grades	0.023	0.040**
	(0.015)	(0.019)
1=Parent knows grades in school	0.010	-0.017
	(0.018)	(0.025)
1=Parent willing to help	-0.012	0.058***
	(0.011)	(0.022)
PSU test scores	-0.014	0.111**
	(0.031)	(0.053)
Observations	71,769	27,810

Notes: Table shows the effect of sibling school starting age on parental investments in grades 4 and 8 for younger siblings (older-to-younger) – Column (1) – and older siblings (younger-to-older) – Column (2). Results are based on the empirical strategy that implements a RD following the methodology proposed by Calonico et al. (2014), with a polynomial of order one and weighted by triangular kernel. Standard errors are clustered at the running variable at daily level. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

already struggling at school, parents might decide to avoid or reduce any resource reallocation from their older child to the younger. This could explain why the positive spillover effect is only present for older children who exhibit higher levels of school performance.

Finally, the results might also be explained by the differences in the size of the "shock" for older and younger siblings. In particular, we would expect that a smaller "shock" has smaller own effects and smaller spillovers within the family. Older siblings are less affected by the school starting age, and so the spillover effects produced by the shock could also be lower.

3.6 Conclusions

This chapter asks whether children's school starting age affect long run educational outcomes of their siblings. To address this question, I combine administrative data from Chile and take advantage of the variation generated by an enrolment cutoff. I start by documenting the effect of own school starting age on student achievement. The findings are consistent with the literature, older school starters perform better on in-school outcomes, such as standardised test and school grades. Moreover, these effects persist even 12-13 years after starting the school: children who start the school later tend to score higher in college admission exams and are more likely to enrol in college. I also show that older siblings are less affected by school entry policies vis-à-vis younger siblings.

Then, I test whether school entry policies have spillover effects within the family. In this regard, I find positive spillovers from younger-to-older siblings. Specifically, older siblings score 0.05σ higher in college admission exams if their sibling enters school at an older age. No spillover effect is found from older-to-younger siblings. In addition, the gains accrue only to students coming from high-income families, who are close in age to the younger sibling and have higher school grades before their sibling make the transition into school. Altogether, the findings of this chapter are consistent with a model of joint formation of human capital, and suggest the need to account for a wider set of spillover effects when designing, implementing, and evaluating educational interventions.

Moreover, by exploiting rich information from a survey of parents and students, I explore several potential channels through which these results can be explained. I find that older siblings tend to receive less attention from their parents when their sibling starts school at a younger age. I interpret this finding as evidence of parents responding to endowment differentials among their children following a compensatory strategy, i.e. devoting more resources to children with lower endowments. However, these results are far from conclusive. The questions remain as to which mechanisms can explain why the younger sibling's school starting age can have an effect on older sibling's learning and performance; and why we do not observe an effect in the other direction, i.e. spillovers from older-to-younger. This is left for future research.

Bibliography

Acemoglu, D. and Autor, D. (2012), 'What Does Human Capital Do? A Review of Goldin and Katz's The Race between Education and Technology', *Journal of Economic Literature* **50**(2), 426–463.

Aguirre, J. and Matta, J. (2021), 'Walking in your footsteps: Sibling spillovers in higher education choices', *Economics of Education Review* **80**, 102062.

Aizer, A. and Cunha, F. (2012), The Production of Human Capital: Endowments, Investments and Fertility, Working Paper 18429, National Bureau of Economic Research.

Altmejd, A., Barrios-Fernández, A., Drlje, M., Goodman, J., Hurwitz, M., Kovac, D., Mulhern, C., Neilson, C. and Smith, J. (2021), 'O Brother, Where Start Thou? Sibling Spillovers on College and Major Choice in Four Countries', *The Quarterly Journal of Economics* **136**(3), 1831–1886.

Arrow, K. (1973), *Discrimination in Labor Markets*, Princeton University Press.

Attar, I. and Cohen-Zada, D. (2018), 'The effect of school entrance age on educational outcomes: Evidence using multiple cutoff dates and exact date of birth', *Journal of Economic Behavior & Organization* **153**, 38–57.

Aucejo, E. M., Coate, P., Fruehwirth, J. C., Kelly, S. and Mozenter, Z. (2018), Teacher effectiveness and classroom composition. CEP Discussion Paper 1574, Centre for Economic Performance, London School of Economics and Political Science.

Azmat, G., Calsamiglia, C. and Iriberri, N. (2016), 'Gender Differences in Response to Big Stakes', *Journal of the European Economic Association* **14**(6), 1372–1400.

Barclay, K. J. (2018), 'The birth order paradox: Sibling differences in educational attainment', *Research in Social Stratification and Mobility* **54**, 56–65.

Becker, G. S. and Tomes, N. (1976), 'Child Endowments and the Quantity and Quality of Children', *Journal of Political Economy* **84**(4), S143–S162.

Bedard, K. and Dhuey, E. (2006), 'The Persistence of Early Childhood Maturity: International Evidence of Long-Run Age Effects', *The Quarterly Journal of Economics* **121**(4), 1437–1472.

Behrman, J. R., Pollak, R. A. and Taubman, P. (1982), 'Parental Preferences and Provision for Progeny', *Journal of Political Economy* **90**(1), 52–73.

Behrman, J. R., Tincani, M. M., Todd, P. E. and Wolpin, K. I. (2016), 'Teacher Quality in Public and Private Schools under a Voucher System: The Case of Chile', *Journal of Labor Economics* **34**(2), 319–362.

Bellei, C., Valenzuela, J. and De los Ríos, D. (2010), Segregación Escolar en Chile, *in* S. Martinic and G. Elacqua, eds, '¿Fin de Ciclo: Cambios en la Gobernanza del Sistema Educativo?', Pontificia Universidad Católica de Chile & UNESCO, chapter 8, pp. 209–229.

Bertrand, M. and Pan, J. (2013), 'The Trouble with Boys: Social Influences and the Gender Gap in Disruptive Behavior', *American Economic Journal: Applied Economics* **5**(1), 32–64.

Bertrand, M. and Schoar, A. (2003), 'Managing with Style: The Effect of Managers on Firm Policies', *The Quarterly Journal of Economics* **118**(4), 1169–1208.

Bettinger, E. P. and Long, B. T. (2005), 'Do Faculty Serve as Role Models? The Impact of Instructor Gender on Female Students', *American Economic Review* **95**(2), 152–157.

Bharadwaj, P., Eberhard, J. P. and Neilson, C. A. (2018), 'Health at Birth, Parental Investments, and Academic Outcomes', *Journal of Labor Economics* **36**(2), 349–394.

Bietenbeck, J. (2014), ‘Teaching practices and cognitive skills’, *Labour Economics* **30**, 143–153.

Bietenbeck, J., Piopiunik, M. and Wiederhold, S. (2018), ‘Africa’s Skill Tragedy: Does Teachers’ Lack of Knowledge Lead to Low Student Performance?’, *Journal of Human Resources* **53**(3), 553–578.

Black, S. (2000), ‘Together Again: The Practice of Looping Keeps Students with the Same Teachers’, *American School Board Journal* **187**(6), 40–43.

Black, S. E., Devereux, P. J. and Salvanes, K. G. (2005), ‘The More the Merrier? The Effect of Family Size and Birth Order on Children’s Education’, *The Quarterly Journal of Economics* **120**(2), 669–700.

Black, S. E., Devereux, P. J. and Salvanes, K. G. (2011), ‘Too Young to Leave the Nest? The Effects of School Starting Age’, *The Review of Economics and Statistics* **93**(2), 455–467.

Blank, R. (1991), ‘The Effects of Double-Blind versus Single-Blind Reviewing: Experimental Evidence from The American Economic Review’, *American Economic Review* **81**(5), 1041–1067.

Blume, L. E., Brock, W. A., Durlauf, S. N. and Ioannides, Y. M. (2011), Identification of Social Interactions, Vol. 1 of *Handbook of Social Economics*, North-Holland, chapter 18, pp. 853–964.

Bogart, V. S. (2002), The Effects of Looping on the Academic Achievement of Elementary School Students, PhD thesis, East Tennessee State University.

Bonesrønning, H. (2008), ‘The Effect of Grading Practices on Gender Differences in Academic Performance’, *Bulletin of Economic Research* **60**(3), 245–264.

Botelho, F., Madeira, R. A. and Rangel, M. A. (2015), ‘Racial Discrimination in Grading: Evidence from Brazil’, *American Economic Journal: Applied Economics* **7**(4), 37–52.

Breen, R. (2019), ‘Education and intergenerational social mobility in the US and four European countries’, *Oxford Review of Economic Policy* **35**(3), 445–466.

Brookhart, S. M., Guskey, T. R., Bowers, A. J., McMillan, J. H., Smith, J. K., Smith, L. F., Stevens, M. T. and Welsh, M. E. (2016), 'A Century of Grading Research: Meaning and Value in the Most Common Educational Measure', *Review of Educational Research* **86**(4), 803–848.

Bryk, A. S., Sebring, P. B., Allensworth, E., Easton, J. Q. and Luppescu, S. (2010), *Organizing schools for improvement: Lessons from Chicago*, University of Chicago Press.

Burgess, S. and Greaves, E. (2013), 'Test Scores, Subjective Assessment, and Stereotyping of Ethnic Minorities', *Journal of Labor Economics* **31**(3), 535–576.

Burke, D. L. (1996), 'Multi-year teacher/student relationships are a long-overdue arrangement', *Phi Delta Kappan* **77**(5), 360.

Calonico, S., Cattaneo, M. D. and Titiunik, R. (2014), 'Robust nonparametric confidence intervals for regression-discontinuity designs', *Econometrica* **82**(6), 2295–2326.

Casula, L. and Liberto, A. D. (2017), Teacher assessments versus standardized tests: is acting "girly" an advantage?, Working Paper CRENoS 201701, Centre for North South Economic Research, University of Cagliari and Sassari, Sardinia.

Cattaneo, M. D., Jansson, M. and Ma, X. (2018), 'Manipulation Testing Based on Density Discontinuity', *The Stata Journal* **18**(1), 234–261.

Ceci, S. J., Ginther, D. K., Kahn, S. and Williams, W. M. (2014), 'Women in Academic Science: A Changing Landscape', *Psychological Science in the Public Interest* **15**(3), 75–141.

Chetty, R., Friedman, J. N. and Rockoff, J. E. (2014*a*), 'Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates', *American Economic Review* **104**(9), 2593–2632.

Chetty, R., Friedman, J. N. and Rockoff, J. E. (2014*b*), 'Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood', *American Economic Review* **104**(9), 2633–2679.

Chetty, R., Friedman, J. N., Saez, E., Turner, N. and Yagan, D. (2020), ‘Income Segregation and Intergenerational Mobility Across Colleges in the United States’, *The Quarterly Journal of Economics* **135**(3), 1567–1633.

Cistone, P. and Shneyderman, A. (2004), ‘Looping: An Empirical Evaluation’, *International Journal of Educational Policy, Research, and Practice: Reconceptualizing Childhood Studies* **5**(1), 47–61.

Clotfelter, C. T., Ladd, H. F. and Vigdor, J. (2005), ‘Who teaches whom? Race and the distribution of novice teachers’, *Economics of Education review* **24**(4), 377–392.

Clotfelter, C. T., Ladd, H. F. and Vigdor, J. L. (2006), ‘Teacher-student matching and the assessment of teacher effectiveness’, *Journal of Human Resources* **41**(4), 778–820.

Clotfelter, C. T., Ladd, H. F. and Vigdor, J. L. (2010), ‘Teacher credentials and student achievement in high school a cross-subject analysis with student fixed effects’, *Journal of Human Resources* **45**(3), 655–681.

Cohen-Vogel, L. (2011), ‘“Staffing to the test”: are today’s school personnel practices evidence based?’, *Educational Evaluation and Policy Analysis* **33**(4), 483–505.

Comi, S. L., Argentin, G., Gui, M., Origo, F. and Pagani, L. (2017), ‘Is it the way they use it? Teachers, ICT and student achievement’, *Economics of Education Review* **56**, 24–39.

Cook, P. J. and Kang, S. (2016), ‘Birthdays, Schooling, and Crime: Regression-Discontinuity Analysis of School Performance, Delinquency, Dropout, and Crime Initiation’, *American Economic Journal: Applied Economics* **8**(1), 33–57.

Cooper, H. and Good, T. (1983), *Pygmalion grows up: Studies in the expectation communication process*, New York: Longman.

Cornwell, C., Mustard, D. B. and Parys, J. V. (2013), ‘Non-Cognitive Skills and the Gender Disparities in Test Scores and Teacher Assessments: Evidence from Primary School’, *Journal of Human Resources* **48**(1), 236–264.

Correa, J. A., Parro, F. and Reyes, L. (2015), 'Self-selection in the market of teachers', *Applied Economics* **47**(13), 1331–1349.

Correia, S. (2016), A Feasible Estimator for Linear Models with Multi-Way Fixed Effects. Mimeo, Duke University.

Crawford, C., Dearden, L. and Meghir, C. (2007), When You Are Born Matters: The Imapct of Date of Birth on Child Cognitive Outcomes in England, CEE Discussion Papers 93, Centre for the Economics of Education, LSE.

Cuesta, J. I., González, F. and Philippi, C. L. (2020), 'Distorted quality signals in school markets', *Journal of Development Economics* **147**, 102532.

Cunha, F. and Heckman, J. (2008), 'Formulating, identifying and estimating the technology of cognitive and noncognitive skill formation', *Journal of Human Resources* **43**(4), 738–782.

Currie, J. and Almond, D. (2011), Human capital development before age five, Vol. 4 of *Handbook of Labor Economics*, Elsevier, pp. 1315–1486.

Dahl, G. B., Rooth, D.-O. and Stenberg, A. (2020), Family Spillovers in Field of Study, Working Paper 27618, National Bureau of Economic Research.

Datar, A., Kilburn, M. and Loughran, D. (2010), 'Endowments and parental investments in infancy and early childhood', *Demography* **47**(1), 145–162.

Dee, T. (2005), 'A Teacher like Me: Does Race, Ethnicity, or Gender Matter?', *The American Economic Review* **95**(2), 158–165.

Dee, T. S. (2007), 'Teachers and the gender gaps in student achievement', *Journal of Human Resources* **42**(3), 528–554.

Del Bono, E., Ermisch, J. and Francesconi, M. (2012), 'Intrafamily Resource Allocations: A Dynamic Structural Model of Birth Weight', *Journal of Labor Economics* **30**(3), 65–706.

Depew, B. and Eren, O. (2016), 'Born on the wrong day? School entry age and juvenile crime', *Journal of Urban Economics* **96**, 73–90.

Dhuey, E., Figlio, D., Karbownik, K. and Roth, J. (2019), 'School Starting Age and Cognitive Development', *Journal of Policy Analysis and Management* **38**(3), 538–578.

Dobkin, C. and Ferreira, F. (2010), 'Do school entry laws affect educational attainment and labor market outcomes?', *Economics of Education Review* **29**(1), 40–54.

Duckworth, A. and Seligman, M. (2006), 'Self-discipline gives girls the edge: Gender in self-discipline, grades, and achievement test scores', *Journal of Educational Psychology* **98**, 198–208.

Dunn, J. (2007), Siblings and socialization, *in* J. E. Grusec and P. D. Hastings, eds, 'Handbook of socialization: Theory and research', New York: The Guilford Press, pp. 309–327.

Dustan, A. (2018), 'Family networks and school choice', *Journal of Development Economics* **134**, 372–391.

Epple, D. and Romano, R. E. (2011), Chapter 20 - Peer Effects in Education: A Survey of the Theory and Evidence, Vol. 1 of *Handbook of Social Economics*, North-Holland, pp. 1053–1163.

Falch, T. and Naper, L. R. (2013), 'Educational evaluation schemes and gender gaps in student achievement', *Economics of Education Review* **36**, 12–25.

Feng, L. (2010), 'Hire today, gone tomorrow: new teacher classroom assignments and teacher mobility', *Education Finance and Policy* **5**(3), 278–316.

Ffrench-Davis, R. (2010), *Economic Reforms in Chile: From Dictatorship to Democracy*, Palgrave Macmillan UK.

Figlio, D. N. and Lucas, M. E. (2004), 'Do high grading standards affect student performance?', *Journal of Public Economics* **88**(9-10), 1815–1834.

Fitzpatrick, M. D. and Lovenheim, M. F. (2014), 'Early retirement incentives and student achievement', *American Economic Journal: Economic Policy* **6**(3), 120–54.

Franz, D. P., Thompson, N. L., Fuller, B., Hare, R. D., Miller, N. C. and Walker, J. (2010), 'Evaluating mathematics achievement of middle school students in a looping environment', *School Science and Mathematics* **110**(6), 298–308.

Fredriksson, P. and Öckert, B. (2013), 'Life-cycle Effects of Age at School Start', *The Economic Journal* **124**(579), 977–1004.

Frijters, P., Johnston, D. W., Shah, M. and Shields, M. A. (2013), 'Intrahousehold Resource Allocation: Do Parents Reduce or Reinforce Child Ability Gaps?', *Demography* **50**(6), 2187–2208.

Fryer, R. G. J. (2018), 'The "pupil" factory: specialization and the production of human capital in schools', *American Economic Review* **108**(3), 616–656.

Gallegos, S. and Celhay, P. (2020), Early Skill Effects on Types of Parental Investments and Long-Run Outcomes, Working Papers 2020-014, Human Capital and Economic Opportunity Working Group.

Gneezy, U., Niederle, M. and Rustichini, A. (2003), 'Performance in competitive environments: Gender differences', *Quarterly Journal of Economics* **118**(3), 1049–1074.

Gneezy, U. and Rustichini, A. (2004), 'Gender and Competition at a Young Age', *American Economic Review* **94**(2), 377–381.

Goldin, C. and Katz, L. F. (2008), *The Race between Education and Technology*, Harvard University Press.

Goldin, C. and Rouse, C. (2000), 'Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians', *American Economic Review* **90**(4), 715–741.

Graham, B. S., Ridder, G., Thiemann, P. M. and Zamarro, G. (2020), Teacher-to-Classroom Assignment and Student Achievement. NBER Working Paper No. 27543.

Guimaraes, P. and Portugal, P. (2010), 'A simple feasible procedure to fit models with high-dimensional fixed effects', *The Stata Journal* **10**(4), 628–649.

Hahn, J., Todd, P. and der Klaauw, W. V. (2001), 'Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design', *Econometrica* **69**(1), 201–209.

Hanushek, E. A., Kain, J. F. and Rivkin, S. G. (2004), 'Why public schools lose teachers', *Journal of Human Resources* **39**(2), 326–354.

Hanushek, E. A. and Woessmann, L. (2008), 'The Role of Cognitive Skills in Economic Development', *Journal of Economic Literature* **46**(3), 607–668.

Harris, D. N. and Sass, T. R. (2011), 'Teacher training, teacher quality and student achievement', *Journal of Public Economics* **95**(7), 798–812.

Hill, A. J. and Jones, D. B. (2018), 'A teacher who knows me: The academic benefits of repeat student-teacher matches', *Economics of Education Review* **64**, 1–12.

Hinnerich, B. T., Höglin, E. and Johannesson, M. (2011), 'Are boys discriminated in Swedish high schools?', *Economics of Education Review* **30**(4), 682–690.

Hoffmann, F. and Oreopoulos, P. (2009), 'A professor like me: the influence of instructor gender on college achievement', *Journal of Human Resources* **44**(2), 479–494.

Holland, P. W. (1986), 'Statistics and Causal Inference', *Journal of the American Statistical Association* **81**(396), 945–960.

Hsieh, C.-T. and Urquiola, M. (2006), 'The effects of generalized school choice on achievement and stratification: Evidence from Chile's voucher program', *Journal of Public Economics* **90**(8), 1477–1503.

Imbens, G. W. and Lemieux, T. (2008), 'Regression discontinuity designs: A guide to practice', *Journal of Econometrics* **142**(2), 615–635.

Joensen, J. S. and Nielsen, H. S. (2018), 'Spillovers in education choice', *Journal of Public Economics* **157**, 158–183.

Kalogrides, D., Loeb, S. and Béteille, T. (2013), 'Systematic sorting: Teacher characteristics and class assignments', *Sociology of Education* **86**(2), 103–123.

Karbownik, K. and Özek, U. (2019), Setting a Good Example? Examining Sibling Spillovers in Educational Achievement Using a Regression Discontinuity Design, Working Paper 26411, National Bureau of Economic Research.

Kenney-Benson, G., Pomerantz, E., Ryan, A. and Patrick, H. (2006), 'Sex differences in math performance: The role of children's approach to schoolwork', *Developmental Psychology* **42**(1), 11–26.

Kerr, D. L. (2002), "In the loop": responses about looping at the middle school level as seen through different lenses, PhD thesis, National-Louis University.

Klugman, J. (2017), 'Essential or expendable supports? Assessing the relationship between school climate and student outcomes', *Sociological Science* **4**, 31–53.

Kraft, M. A., Marinell, W. H. and Shen-Wei Yee, D. (2016), 'School organizational contexts, teacher turnover, and student achievement: evidence from panel data', *American Educational Research Journal* **53**(5), 1411–1449.

Landersø, R. K., Nielsen, H. S. and Simonsen, M. (2020), 'Effects of School Starting Age on the Family', *Journal of Human Resources* **55**(4), 1258–1286.

Landersø, R., Nielsen, H. S. and Simonsen, M. (2017), 'School Starting Age and the Crime-age Profile', *The Economic Journal* **127**(602), 1096–1118.

Lavy, V. (2008), 'Do gender stereotypes reduce girls' or boys' human capital outcomes? Evidence from a natural experiment', *Journal of Public Economics* **92**(10-11), 2083–2105.

Lavy, V. (2015), 'Do differences in schools' instruction time explain international achievement gaps? Evidence from developed and developing countries', *The Economic Journal* **125**(588), F397–F424.

Lavy, V. and Megalokonomou, R. (2019), Persistency in teachers' grading bias and effects on long-term outcomes: university admissions exams and choice of field of study, Working Paper 26021, National Bureau of Economic Research.

Lavy, V. and Sand, E. (2018), ‘On the origins of gender gaps in human capital: Short- and long-term consequences of teachers’ biases’, *Journal of Public Economics* **67**, 263–279.

Lee, D. S. and Lemieux, T. (2010), ‘Regression Discontinuity Designs in Economics’, *Journal of Economic Literature* **48**(2), 281–355.

Little, T. S. and Dacus, N. B. (1999), ‘Looping: Moving up with the class’, *Educational Leadership* **57**(1), 42–45.

Liu, J.-Q. (1997), ‘The emotional bond between teachers and students: Multi-year relationships’, *Phi Delta Kappan* **79**(2), 156.

Lubotsky, D. and Kaestner, R. (2016), ‘Do ‘Skills Beget Skills’? Evidence on the effect of kindergarten entrance age on the evolution of cognitive and non-cognitive skill gaps in childhood’, *Economics of Education Review* **53**, 194–206.

Manski, C. (1993), ‘Identification of Endogenous Social Effects: The Reflection Problem’, *Review of Economic Studies* **60**(3), 531–542.

Matthews, J., Cameron, C. and Morrison, F. (2009), ‘Early Gender Differences in Self-Regulation and Academic Achievement’, *Journal of Educational Psychology* **101**(3), 689–704.

McEwan, P. J. and Shapiro, J. S. (2008), ‘The Benefits of Delayed Primary School Enrollment: Discontinuity Estimates Using Exact Birth Dates’, *The Journal of Human Resources* **43**(1), 1–29.

McHale, S. M., Updegraff, K. A. and Whiteman, S. D. (2012), ‘Sibling Relationships and Influences in Childhood and Adolescence’, *Journal of marriage and the family* **74**(5), 913–930.

Mechtenberg, L. (2009), ‘Cheap Talk in the Classroom: How Biased Grading at School Explains Gender Differences in Achievements, Career Choices and Wages’, *Review of Economic Studies* **76**(4), 1431–1459.

Ministerio de Desarrollo Social (2017), CASEN 2017: Educación, Encuesta de Caracterización Socioeconómica Nacional, Technical report.

Ministerio de Desarrollo Social (2020), Evolución de la pobreza 1990-2017: ¿Cómo ha cambiado Chile?, Technical report.

Mizala, A. and Torche, F. (2012), 'Bringing the schools back in: the stratification of educational achievement in the Chilean voucher system', *International Journal of Educational Development* **32**(1), 132–144.

Nam, K. (2014), 'Until when does the effect of age on academic achievement persist? Evidence from Korean data', *Economics of Education Review* **40**, 106–122.

Nichols, J. D. and Nichols, G. W. (2002), 'The impact of looping and non-looping classroom environments on parental attitudes', *Educational Research Quarterly* **26**(1), 23.

Niederle, M. and Vesterlund, L. (2007), 'Do Women Shy away from Competition? Do Men Compete too Much?', *Quarterly Journal of Economics* **122**(3), 1067–1101.

OECD (2017), *Education in Chile*, OECD Publishing, Paris.

OECD (2018), *OECD Economic Surveys: Chile 2018*, OECD Publishing, Paris.

OECD (2019*a*), *PISA 2018 Results (Volume I): What Students Know and Can Do*, OECD Publishing, Paris.

OECD (2019*b*), *PISA 2018 Results (Volume II): Where All Students Can Succeed*, OECD Publishing, Paris.

Osborne-Lampkin, L. and Cohen-Vogel, L. (2014), '"Spreading the wealth": how principals use performance data to populate classrooms', *Leadership and Policy in Schools* **13**(2), 188–208.

Ost, B. (2014), 'How do teachers improve? The relative importance of specific and general human capital', *American Economic Journal: Applied Economics* **6**(2), 127–51.

Paredes, V. (2014), 'A teacher like me or a student like me? Role model versus teacher bias effect', *Economics of Education Review* **39**, 38–49.

Peña, P. A. (2017), ‘Creating winners and losers: Date of birth, relative age in school, and outcomes in childhood and adulthood’, *Economics of Education Review* **56**, 152–176.

Phelps, E. S. (1972), ‘The Statistical Theory of Racism and Sexism’, *American Economic Review* **62**(4), 659–661.

Protivínský, T. and Münich, D. (2018), ‘Gender Bias in teachers’ grading: What is in the grade’, *Studies in Educational Evaluation* **59**, 141–149.

Puhani, P. and Weber, A. (2007), ‘Does the early bird catch the worm?’, *Empirical Economics* **32**, 359–386.

Ready, D. D., LoGerfo, L. F., Burkam, D. T. and Lee, V. E. (2005), ‘Explaining Girls’ Advantage in Kindergarten Literacy Learning: Do Classroom Behaviors Make a Difference?’, *Elementary School Journal* **106**(1), 21–38.

Restrepo, B. J. (2016), ‘Parental investment responses to a low birth weight outcome: who compensates and who reinforces?’, *Journal of Population Economics* **29**(4), 969–989.

Rivkin, S. G., Hanushek, E. A. and Kain, J. F. (2005), ‘Teachers, Schools, and Academic Achievement’, *Econometrica* **73**(2), 417–458.

Rockoff, J. E. (2004), ‘The impact of individual teachers on student achievement: evidence from panel data’, *American Economic Review* **94**(2), 247–252.

Rosales-Rueda, M. F. (2014), ‘Family investment responses to childhood health conditions: Intrafamily allocation of resources’, *Journal of Health Economics* **37**, 41–57.

Rosenzweig, M. R. and Zhang, J. (2009), ‘Do Population Control Policies Induce More Human Capital Investment? Twins, Birth Weight and China’s “One-Child” Policy’, *The Review of Economic Studies* **76**(3), 1149–1174.

Royer, H. (2009), ‘Separated at Girth: US Twin Estimates of the Effects of Birth Weight’, *American Economic Journal: Applied Economics* **1**(1), 49–85.

Rubin, D. (1974), 'Estimating causal effects of treatments in randomized and nonrandomized studies', *Journal of Educational Psychology* **66**(5), 688–701.

Sacerdote, B. (2011), Peer Effects in Education: How Might They Work, How Big Are They and How Much Do We Know Thus Far?, *in* E. Hanushek, S. Machin and L. Woessmann, eds, 'Handbook of the Economics of Education', Vol. 3 of *Handbook of the Economics of Education*, Elsevier, chapter 4, pp. 249–277.

Sadker, M. and Sadker, D. (1994), *Failing at Fairness: How Our Schools Cheat Girls*, New York: Touchstone.

Salisbury, J., Rees, G. and Gorard, S. (1999), 'Accounting for the Differential Attainment of Boys and Girls at School', *School Leadership and Management* **19**(4), 403–426.

Santiago, P., Fiszbein, A., Jaramillo, S. G. and Radinger, T. (2017), *OECD Reviews of School Resources: Chile 2017*, OECD publishing.

Sanz-de Galdeano, A. and Terskaya, A. (2019), Sibling Differences in Educational Polygenic Scores: How Do Parents React?, IZA Discussion Papers 12375, Institute of Labor Economics (IZA).

Schmidt-Hebbel, K. (2006), 'Chile's Economic Growth', *Cuadernos de Economía* **43**(127), 5–48.

Shurchkov, O. (2012), 'Under Pressure: Gender Differences in Output Quality and Quantity Under Competition and Time Constraints', *Journal of the European Economic Association* **10**(5), 1189–1213.

Terrier, C. (2020), 'Boys lag behind: How teachers' gender biases affect student achievement', *Economics of Education Review* **77**, 101981.

Thapa, A., Cohen, J., Guffey, S. and Higgins-D'Alessandro, A. (2013), 'A review of school climate research', *Review of Educational Research* **83**(3), 357–385.

Tiedemann, J. (2000), 'Gender related beliefs of teachers in elementary school mathematics', *Educational Studies in Mathematics* **41**(2), 191–207.

Todd, P. E. and Wolpin, K. I. (2007), 'The Production of Cognitive Achievement in Children: Home, School, and Racial Test Score Gaps', *Journal of Human Capital* **1**(1), 91–136.

Tourigny, R., Plante, I. and Raby, C. (2019), 'Do students in a looping classroom get higher grades and report a better teacher-student relationship than those in a traditional setting?', *Educational Studies* pp. 1–16.

Tucker, S. C. (2006), The Impact of Looping On Student Achievement On the Colorado Student Assessment Program, Master's thesis, Regis University.

UNESCO (2015), Informe de Resultados TERCE: Tercero Estudio Regional Comparativo y Explicativo, Factores Associados, Technical report, UNESCO's Regional Office for Education in Latin America and the Caribbean OREALC/ UNESCO, Santiago de Chile.

Whiteman, S. D., McHale, S. M. and Soli, A. (2011), 'Theoretical Perspectives on Sibling Relationships', *Journal of family theory & review* **3**(2), 124–139.

World Bank (2021), 'World development indicators'. data retrieved from World Development Indicators, https://databank.worldbank.org/source/world-development-indicators.

Yi, J., Heckman, J. J., Zhang, J. and Conti, G. (2015), 'Early Health Shocks, Intra-household Resource Allocation and Child Outcomes', *The Economic Journal* **125**(588), F347–F371.

Zahorik, J. A. and Dichanz, H. (1994), 'Teaching for Understanding in German Schools', *Educational Leadership* **51**(5), 75–77.

Appendix A

A.1 Gender grading gaps by year-grade combinations

Table A.1. Gender gap in grading using different year-grade combinations: Spanish

Year	4th grade (1)	6th grade (2)	8th grade (3)	10th grade (4)
2011	0.047*** (0.006)	-	0.177*** (0.007)	-
2012	0.054*** (0.006)	-	-	0.180*** (0.011)
2013	0.060*** (0.005)	0.209*** (0.006)	0.241*** (0.007)	0.194*** (0.010)
2014	0.075*** (0.005)	0.175*** (0.006)	0.212*** (0.007)	0.181*** (0.011)
2015	0.077*** (0.005)	0.196*** (0.006)	0.214*** (0.007)	0.110*** (0.011)
2016	0.118*** (0.006)	0.209*** (0.006)	-	0.118*** (0.011)
2017	0.120*** (0.006)	-	0.200*** (0.007)	0.125*** (0.011)
2018	0.126*** (0.006)	0.188*** (0.007)	-	0.121*** (0.011)

Notes: Each cell in this table reports the gender gap coefficient (γ) from Equation 1.7 for each year-grade combination available. Standard errors are clustered at school level and are reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

TABLE A.2. Gender gap in grading using different year-grade combinations: Math

Year	4th grade (1)	6th grade (2)	8th grade (3)	10th grade (4)
2011	0.102*** (0.005)	-	0.157*** (0.007)	-
2012	0.081*** (0.005)	-	-	0.104*** (0.010)
2013	0.063*** (0.005)	0.189*** (0.006)	0.201*** (0.007)	0.152*** (0.010)
2014	0.075*** (0.005)	0.195*** (0.006)	0.195*** (0.007)	0.110*** (0.011)
2015	0.057*** (0.006)	0.183*** (0.006)	0.254*** (0.007)	0.112*** (0.010)
2016	0.086*** (0.006)	0.207*** (0.006)	-	0.123*** (0.010)
2017	0.102*** (0.006)	-	0.200*** (0.007)	0.143*** (0.009)
2018	0.102*** (0.006)	0.173*** (0.006)	-	0.171*** (0.009)

Notes: Each cell in this table reports the gender gap coefficient (γ) from Equation 1.7 for each year-grade combination available. Standard errors are clustered at school level and are reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

A.2 Gender gap by student's ability

TABLE A.3. Gender gap by student's ability

	All	Low-ability (below median)	High-ability (above median)
	(1)	(2)	(3)
(a) Subject: Spanish			
6th grade	0.161***	0.146***	0.220***
	(0.003)	(0.004)	(0.004)
8th grade	0.175***	0.168***	0.233***
	(0.004)	(0.005)	(0.005)
10th grade	0.118***	0.121***	0.157***
	(0.007)	(0.009)	(0.006)
(b) Subject: Math			
6th grade	0.200***	0.188***	0.194***
	(0.003)	(0.004)	(0.004)
8th grade	0.222***	0.210***	0.199***
	(0.004)	(0.005)	(0.005)
10th grade	0.155***	0.128***	0.143***
	(0.008)	(0.010)	(0.006)

Notes: Each cell in this table reports the gender gap coefficient (γ) from Equation 1.6 for each category. Student's ability is measured using 4th grade SIMCE test scores. Low (high) ability students refers to students below (above) the median ability. Standard errors are clustered at school level and are reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

A.3 Gender grading gaps: additional checks

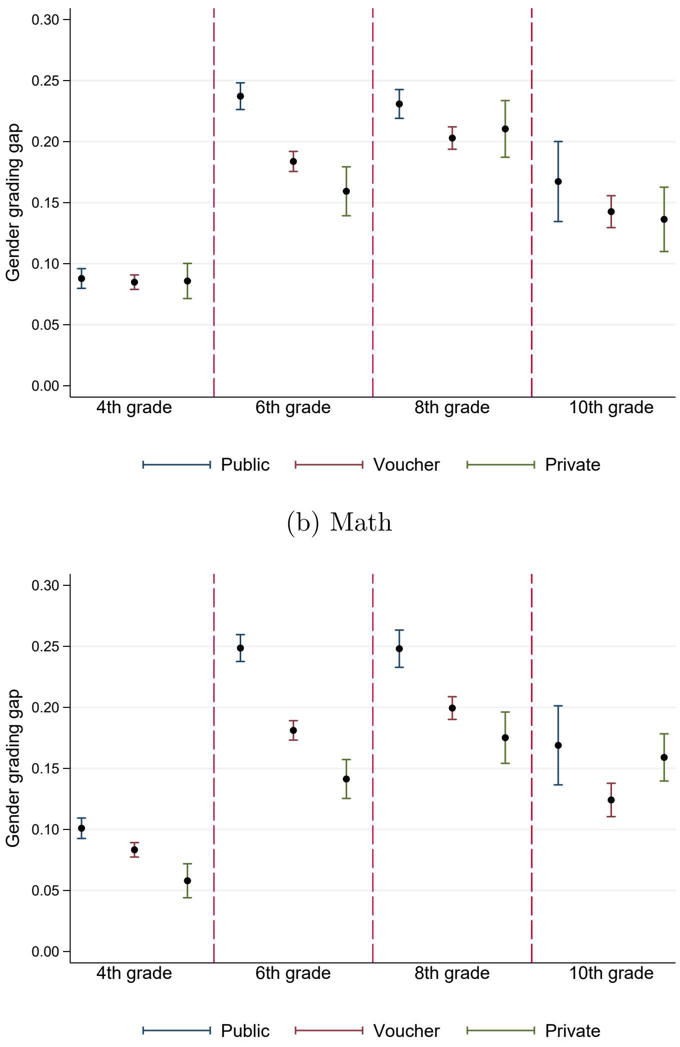
a. By school type:

Table A.4. Gender gap in grading by school type

	All (1)	Public (2)	Voucher (3)	Private (4)
(a) Subject: Spanish				
4th grade	0.083*** (0.002)	0.088*** (0.004)	0.085*** (0.003)	0.086*** (0.007)
6th grade	0.196*** (0.003)	0.237*** (0.006)	0.184*** (0.004)	0.159*** (0.010)
8th grade	0.209*** (0.004)	0.231*** (0.006)	0.203*** (0.005)	0.210*** (0.012)
10th grade	0.147*** (0.007)	0.167*** (0.017)	0.143*** (0.007)	0.136*** (0.013)
(b) Subject: Math				
4th grade	0.083*** (0.002)	0.101*** (0.004)	0.083*** (0.003)	0.058*** (0.007)
6th grade	0.190*** (0.003)	0.249*** (0.006)	0.181*** (0.004)	0.141*** (0.008)
8th grade	0.202*** (0.004)	0.248*** (0.008)	0.199*** (0.005)	0.175*** (0.011)
10th grade	0.131*** (0.007)	0.169*** (0.017)	0.124*** (0.007)	0.159*** (0.010)

Notes: Each cell in this table reports the gender gap coefficient (γ) from Equation 1.7 for each category. Standard errors are clustered at school level and are reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Figure A.1. Gender gap by school type



(a) Spanish

(b) Math

Notes: The graph shows the estimates of γ in Equation 1.7 with 95% confidence interval for each category.

b. By school size:

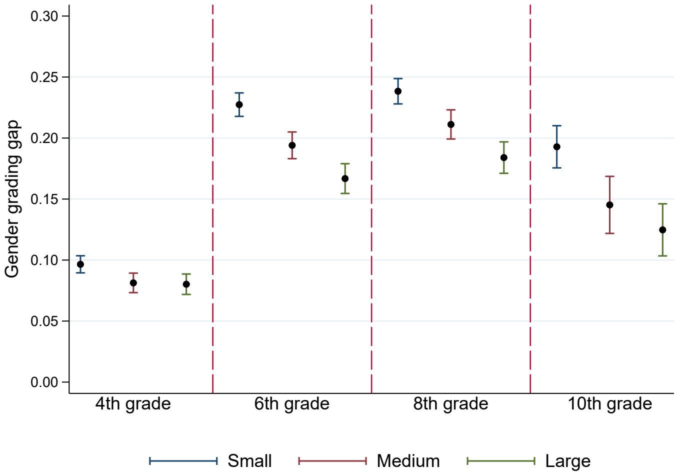
Table A.5. Gender gap in grading by school size

	All (1)	Small (2)	Medium (3)	Large (4)
(a) Subject: Spanish				
4th grade	0.083***	0.097***	0.081***	0.080***
	(0.002)	(0.004)	(0.004)	(0.004)
6th grade	0.196***	0.227***	0.194***	0.167***
	(0.003)	(0.005)	(0.006)	(0.006)
8th grade	0.209***	0.238***	0.211***	0.184***
	(0.004)	(0.005)	(0.006)	(0.007)
10th grade	0.147***	0.193***	0.145***	0.125***
	(0.007)	(0.009)	(0.012)	(0.011)
(b) Subject: Math				
4th grade	0.083***	0.104***	0.085***	0.070***
	(0.002)	(0.004)	(0.004)	(0.004)
6th grade	0.190***	0.225***	0.193***	0.167***
	(0.003)	(0.005)	(0.006)	(0.006)
8th grade	0.202***	0.257***	0.197***	0.173***
	(0.004)	(0.006)	(0.007)	(0.006)
10th grade	0.131***	0.165***	0.136***	0.115***
	(0.007)	(0.009)	(0.012)	(0.011)

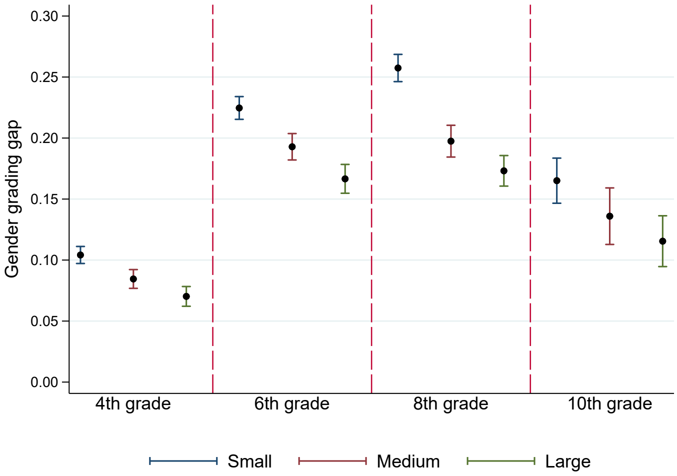
Notes: The sample is divided using an indicator of whether the school is among the lowest third (*Small*), middle third (*Medium*) or highest third (*Large*) of the school size distribution. Each cell in this table reports the gender gap coefficient (γ) from Equation 1.7 for each category. Standard errors are clustered at school level and are reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

FIGURE A.2. Gender gap by school size

(a) Spanish



(b) Math



Notes: The sample is divided using an indicator of whether the school is among the lowest third (*Small*), middle third (*Medium*) or highest third (*Large*) of the school size distribution. The graph shows the estimates of γ in Equation 1.7 with 95% confidence interval for each category.

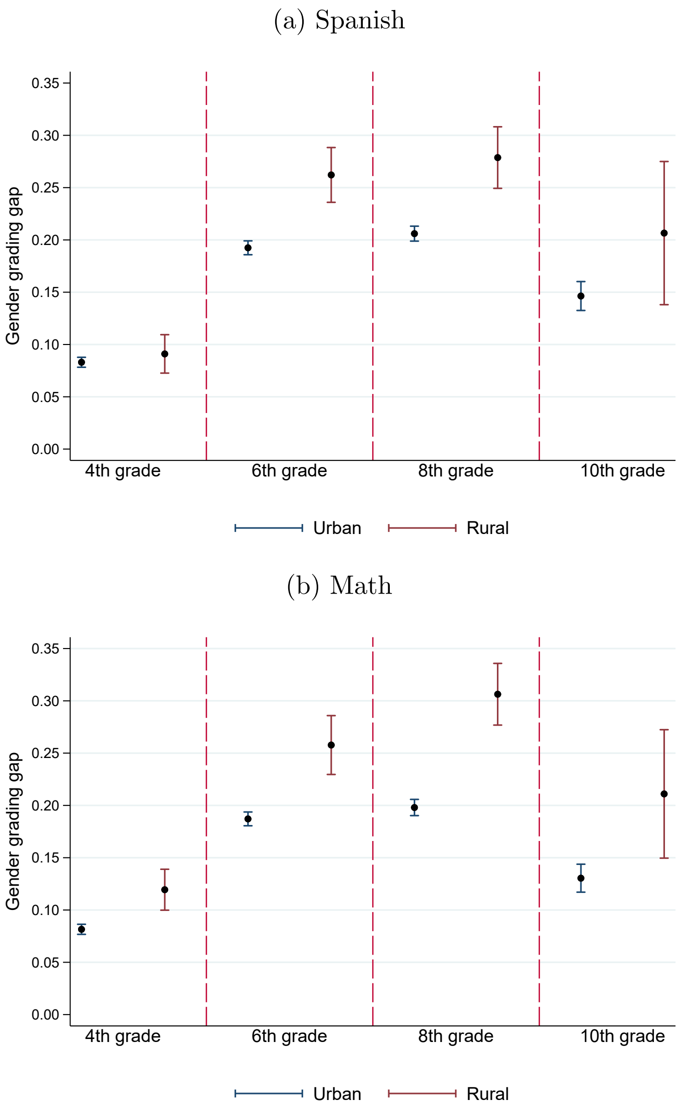
c. Urban and rural schools:

TABLE A.6. Gender gap in grading in urban and rural schools

	All (1)	Urban (2)	Rural (3)
(a) Subject: Spanish			
4th grade	0.083***	0.083***	0.091***
	(0.002)	(0.002)	(0.009)
6th grade	0.196***	0.192***	0.262***
	(0.003)	(0.003)	(0.013)
8th grade	0.209***	0.206***	0.279***
	(0.004)	(0.004)	(0.015)
10th grade	0.147***	0.146***	0.207***
	(0.007)	(0.007)	(0.035)
(b) Subject: Math			
4th grade	0.083***	0.082***	0.119***
	(0.002)	(0.002)	(0.010)
6th grade	0.190***	0.187***	0.258***
	(0.003)	(0.003)	(0.014)
8th grade	0.202***	0.198***	0.306***
	(0.004)	(0.004)	(0.015)
10th grade	0.131***	0.130***	0.211***
	(0.007)	(0.007)	(0.031)

Notes: Each cell in this table reports the gender gap coefficient (γ) from Equation 1.7 for each category. Standard errors are clustered at school level and are reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Figure A.3. Gender gap in urban and rural schools


(a) Spanish

(b) Math

Notes: The graph shows the estimates of γ in Equation 1.7 with 95% confidence interval for each category.

d. By region:

Table A.7. Gender gap in grading by region: Spanish

	All	Region 1	Region 2	Region 3
4th grade	0.083*** (0.002)	0.063*** (0.013)	0.068*** (0.011)	0.057*** (0.018)
6th grade	0.196*** (0.003)	0.176*** (0.018)	0.170*** (0.014)	0.180*** (0.029)
8th grade	0.209*** (0.004)	0.242*** (0.024)	0.212*** (0.016)	0.154*** (0.022)
10th grade	0.147*** (0.007)	0.177*** (0.026)	0.142* (0.073)	0.125* (0.066)

	Region 4	Region 5	Region 6	Region 7
4th grade	0.096*** (0.009)	0.093*** (0.007)	0.125*** (0.010)	0.086*** (0.011)
6th grade	0.207*** (0.014)	0.182*** (0.011)	0.242*** (0.014)	0.225*** (0.015)
8th grade	0.225*** (0.018)	0.208*** (0.012)	0.229*** (0.017)	0.245*** (0.014)
10th grade	0.143*** (0.034)	0.196*** (0.026)	0.250*** (0.036)	0.177*** (0.016)

	Region 8	Region 9	Region 10	Region 11
4th grade	0.094*** (0.006)	0.096*** (0.009)	0.096*** (0.011)	0.124*** (0.023)
6th grade	0.220*** (0.009)	0.239*** (0.013)	0.213*** (0.015)	0.177*** (0.037)
8th grade	0.216*** (0.010)	0.230*** (0.015)	0.230*** (0.015)	0.206*** (0.055)
10th grade	0.146*** (0.017)	0.169*** (0.034)	0.150*** (0.028)	0.147*** (0.026)

	Region 12	Region 13	Region 14	Region 15
4th grade	0.023 (0.030)	0.075*** (0.004)	0.059*** (0.018)	0.012 (0.017)
6th grade	0.154*** (0.038)	0.176*** (0.005)	0.246*** (0.022)	0.167*** (0.030)
8th grade	0.205*** (0.032)	0.191*** (0.005)	0.226*** (0.022)	0.254*** (0.028)
10th grade	0.144*** (0.046)	0.110*** (0.009)	0.163*** (0.034)	0.124*** (0.022)

Notes: Each cell in this table reports the gender gap coefficient (γ) from Equation 1.7 for each region. Standard errors are clustered at school level and are reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table A.8. Gender gap in grading by region: Math

	All	Region 1	Region 2	Region 3
4th grade	0.083***	0.061***	0.098***	0.104***
	(0.002)	(0.017)	(0.011)	(0.018)
6th grade	0.190***	0.170***	0.197***	0.187***
	(0.003)	(0.023)	(0.018)	(0.025)
8th grade	0.202***	0.215***	0.237***	0.189***
	(0.004)	(0.023)	(0.020)	(0.032)
10th grade	0.131***	0.199***	0.109**	0.165**
	(0.007)	(0.020)	(0.053)	(0.062)

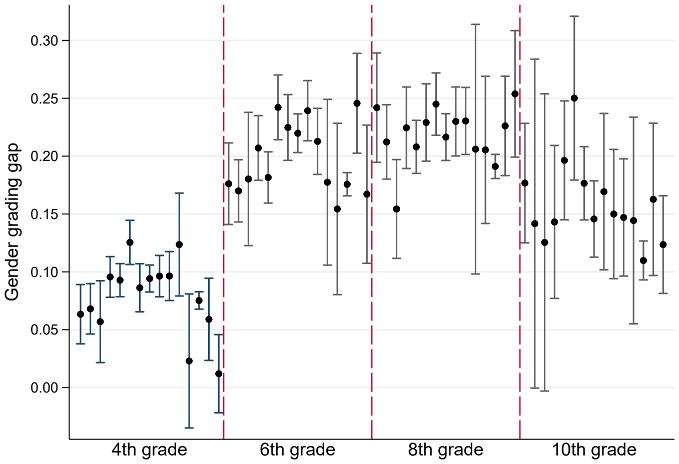
	Region 4	Region 5	Region 6	Region 7
4th grade	0.087***	0.075***	0.112***	0.089***
	(0.011)	(0.008)	(0.011)	(0.010)
6th grade	0.179***	0.185***	0.225***	0.232***
	(0.015)	(0.010)	(0.015)	(0.013)
8th grade	0.204***	0.193***	0.229***	0.226***
	(0.020)	(0.011)	(0.016)	(0.016)
10th grade	0.053	0.180***	0.189***	0.148***
	(0.035)	(0.021)	(0.029)	(0.020)

	Region 8	Region 9	Region 10	Region 11
4th grade	0.096***	0.126***	0.082***	0.135***
	(0.006)	(0.010)	(0.010)	(0.026)
6th grade	0.206***	0.246***	0.209***	0.199***
	(0.008)	(0.015)	(0.015)	(0.024)
8th grade	0.210***	0.212***	0.225***	0.114*
	(0.011)	(0.021)	(0.017)	(0.067)
10th grade	0.128***	0.092***	0.098***	0.126**
	(0.018)	(0.026)	(0.032)	(0.057)

	Region 12	Region 13	Region 14	Region 15
4th grade	0.052*	0.071***	0.077***	0.057***
	(0.028)	(0.004)	(0.019)	(0.019)
6th grade	0.147***	0.168***	0.195***	0.157***
	(0.035)	(0.005)	(0.027)	(0.025)
8th grade	0.173***	0.189***	0.211***	0.199***
	(0.029)	(0.006)	(0.034)	(0.027)
10th grade	0.120**	0.123***	0.172***	0.095***
	(0.045)	(0.010)	(0.032)	(0.021)

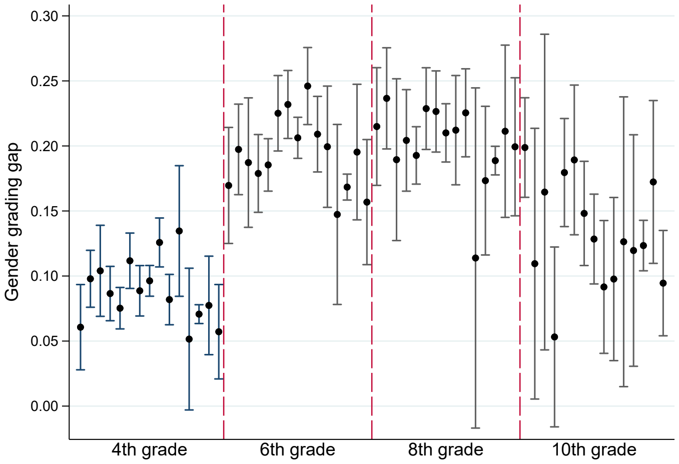
Notes: Each cell in this table reports the gender gap coefficient (γ) from Equation 1.7 for each region. Standard errors are clustered at school level and are reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

FIGURE A.4. Gender gap in grading by region: Spanish



Notes: The graph shows the estimates of γ in Equation 1.7 with 95% confidence interval for each region. Regions are ordered from left to right (*Region 1* to *Region 15*) within each grade.

FIGURE A.5. Gender gap in grading by region: Math



Notes: For each category the graph shows the point estimates with 95% confidence interval. Regions are ordered from left to right (*Region 1* to *Region 15*) within each grade.

A.4 Mean comparison test

Table A.9. Mean comparison test

	Estimation sample	Sample with student behaviour	Difference	Std. err.
(a) Student level				
1=Female	0.50	0.50	-0.001	(0.001)
Father's education	12.11	12.00	0.107***	(0.005)
Mother's education	12.21	12.08	0.130***	(0.004)
Household income	6.14	6.06	0.085***	(0.007)
1=Ethnic group	0.051	0.053	-0.0018***	(0.0003)
1=Foreign student	0.004	0.004	-0.0003***	(0.0001)
4th grade score	0.39	0.42	-0.032***	(0.001)
School attendance	93.31	93.58	-0.271***	(0.006)
1=Grade retention	0.07	0.08	-0.006***	(0.000)
Observations	2,821,911	870,765		
(b) Teacher level				
1=Female	0.76	0.70	0.052***	(0.004)
Experience	13.37	13.30	0.074***	(0.095)
1=Permanent contract	0.57	0.55	0.021***	(0.004)
Working hours	37.44	37.51	-0.063***	(0.050)
Observations	40,044	24,517		
(c) School level				
School enrolment	562.23	587.17	-24.942***	(8.176)
1=Public	0.45	0.45	-0.001***	(0.010)
1=Voucher	0.49	0.49	0.000***	(0.010)
1=Private	0.06	0.06	0.001***	(0.005)
1=Urban	0.89	0.90	-0.015**	(0.006)
1=Metropolitan	0.34	0.34	0.001**	(0.010)
Observations	5,126	4,664		

Notes: Lagged SIMCE test scores are available for 1,890,782 students in the estimation sample. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

A.5 Potential mechanisms

Table A.10. Effects of student behaviour on gender grading gaps, using lagged values

	Spanish			Math		
	(1)	(2)	(3)	(1)	(2)	(3)
Female	0.137***	0.133***	0.003	-0.043***	-0.047***	-0.196***
	(0.003)	(0.003)	(0.049)	(0.003)	(0.003)	(0.045)
Non-blind	0.119***	-1.119***	-1.045***	0.512***	-0.117*	-0.043
	(0.018)	(0.058)	(0.067)	(0.019)	(0.061)	(0.065)
Non-blind × Female	0.181***	0.159***	0.013	0.207***	0.183***	0.040
	(0.005)	(0.005)	(0.067)	(0.005)	(0.005)	(0.063)
Behaviour						
Do homework		0.035***	0.031***		0.055***	0.038***
		(0.004)	(0.005)		(0.003)	(0.004)
Like to study		0.010***	0.015***		0.008**	-0.003
		(0.003)	(0.005)		(0.003)	(0.004)
Grade retention		-0.074***	-0.071***		0.102***	0.095***
		(0.009)	(0.011)		(0.008)	(0.010)
School attendance		0.007***	0.006***		0.013***	0.012***
		(0.000)	(0.000)		(0.000)	(0.000)
Non-blind ×						
Do homework		0.186***	0.181***		0.131***	0.130***
		(0.005)	(0.006)		(0.004)	(0.005)
Like to study		0.065***	0.054***		0.101***	0.096***
		(0.005)	(0.006)		(0.005)	(0.006)
Grade retention		-0.368***	-0.353***		-0.342***	-0.319***
		(0.013)	(0.014)		(0.012)	(0.013)
School attendance		0.012***	0.011***		0.005***	0.004***
		(0.001)	(0.001)		(0.001)	(0.001)
Female ×						
Do homework			0.009			0.038***
			(0.007)			(0.006)
Like to study			-0.010*			0.022***
			(0.006)			(0.005)
Grade retention			-0.010			0.016
			(0.016)			(0.015)
School attendance			0.001***			0.001**
			(0.001)			(0.000)
Female × Non-blind ×						
Do homework			0.011			0.001
			(0.008)			(0.008)
Like to study			0.022***			0.009
			(0.007)			(0.007)
Grade retention			-0.042**			-0.065***
			(0.020)			(0.019)
School attendance			0.001*			0.001**
			(0.001)			(0.001)
R-squared	0.320	0.340	0.340	0.366	0.383	0.383
Observations	621,542	621,542	621,542	621,542	621,542	621,542

Notes: All regressions include student characteristics (father's education, mother's education, household income, indicators of ethnicity and foreign student) and lagged SIMCE test scores, along with their interactions with non-blind test. Standard errors are clustered at school level and are reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table A.11. Effects of student behaviour on gender grading gaps, using raw values

	Spanish			Math		
	(1)	(2)	(3)	(1)	(2)	(3)
Female	0.126***	0.108***	-0.062	-0.035***	-0.051***	-0.315***
	(0.002)	(0.002)	(0.038)	(0.002)	(0.002)	(0.034)
Non-blind	0.122***	-1.684***	-1.583***	0.494***	-0.689***	-0.589***
	(0.015)	(0.059)	(0.064)	(0.017)	(0.065)	(0.065)
Non-blind × Female	0.181***	0.137***	-0.065	0.182***	0.133***	-0.068
	(0.004)	(0.004)	(0.050)	(0.004)	(0.004)	(0.046)
Behaviour						
Do homework		0.060***	0.059***		0.075***	0.069***
		(0.002)	(0.002)		(0.002)	(0.002)
Like to study		0.057***	0.055***		0.032***	0.027***
		(0.001)	(0.002)		(0.001)	(0.002)
Grade retention		-0.017***	-0.015***		0.081***	0.089***
		(0.005)	(0.006)		(0.005)	(0.006)
School attendance		0.008***	0.007***		0.014***	0.013***
		(0.000)	(0.000)		(0.000)	(0.000)
Non-blind ×						
Do homework		0.226***	0.228***		0.184***	0.181***
		(0.003)	(0.003)		(0.002)	(0.003)
Like to study		0.025***	0.015***		0.067***	0.056***
		(0.002)	(0.002)		(0.002)	(0.002)
Grade retention		-0.321***	-0.312***		-0.273***	-0.260***
		(0.007)	(0.008)		(0.008)	(0.008)
School attendance		0.012***	0.011***		0.006***	0.005***
		(0.001)	(0.001)		(0.001)	(0.001)
Female ×						
Do homework			0.003			0.013***
			(0.003)			(0.002)
Like to study			0.003			0.011***
			(0.002)			(0.002)
Grade retention			-0.006			-0.023***
			(0.007)			(0.007)
School attendance			0.002***			0.002***
			(0.000)			(0.000)
Female × **Non-blind** ×						
Do homework			-0.003			0.004
			(0.003)			(0.003)
Like to study			0.021***			0.023***
			(0.003)			(0.003)
Grade retention			-0.024***			-0.034***
			(0.009)			(0.009)
School attendance			0.002***			0.001***
			(0.001)			(0.000)
R-squared	0.321	0.363	0.363	0.363	0.400	0.400
Observations	1,741,530	1,741,530	1,741,530	1,741,530	1,741,530	1,741,530

Notes: All regressions include student characteristics (father's education, mother's education, household income, indicators of ethnicity and foreign student) and lagged SIMCE test scores, along with their interactions with non-blind test. Standard errors are clustered at school level and are reported in parentheses. *$p < 0.10$, **$p < 0.05$, ***$p < 0.01$.

Appendix B

B.1 School panel

In this Appendix we provide evidence on the extent to which looping can be considered a school-level policy. We use data on students between grade 5 and grade 8, during the years 2002-2018. Then we link students with their classroom teacher in four different subjects (Spanish, maths, social sciences and natural sciences). Using this information, we identify repeat matches in grade 6, grade 7 and grade 8. As a result, we obtain a sample of $12,102,819$ students, $8,379$ schools and $444,859$ classes.

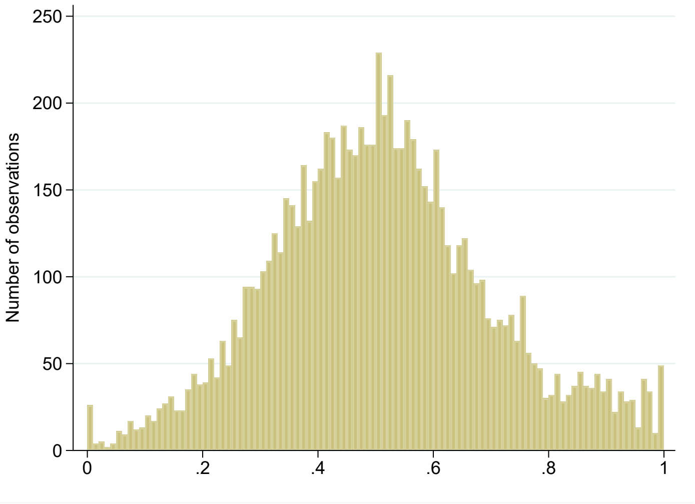
We aggregate this data to the school-year level and calculate the proportion of repeat matches across all subjects and grades, $\bar{R}_{kt}$. At this level, the sample contains $116,812$ observations on $8,379$ schools. A variance decomposition exercise reveals that the variation in looping within schools (overtime) is almost exactly equal to the variation in average looping behaviour between schools. Figure B.1 shows the distribution of $\bar{R}_k$, which indicates that very few schools always or never use repeat matches.

We then aggregate the data to the school-subject-grade-year level, and again compute the proportion of repeat matches, $\bar{R}_{ksgt}$. To quantify how much of the variation in looping can be attributed to schools we estimate the following specification:

$$\bar{R}_{ksgt} = \mu_k + \mu_s + \mu_g + \mu_t \tag{B.1}$$

where μ_k is a school fixed effect, μ_s is a subject fixed effect, μ_g is a grade fixed effect and μ_t is year fixed effect. Column 1 and Column 2 in Table B.1

Figure B.1. Distribution of the proportion of repeat matches at school level



show the benchmark model without and with school fixed effects, respectively. The results show only small variation in repeat matches across subjects and rather larger effects across grades. The inclusion of school fixed effects increases the adjusted R^2 from 2% to only 15%, from which we conclude that the prevalence of looping is only weakly associated with school-level decisions.

Table B.1. Contribution of school fixed effects to the proportion of repeat matches

	(1)	(2)
1=Math	0.023***	0.023***
1=Natural	0.020***	0.020***
1=Social	0.014***	0.014***
1=Grade 7	-0.120***	-0.071***
1=Grade 8	0.018***	0.068***
School FE		Yes
Year FE	Yes	Yes
Adjusted R-squared	0.021	0.147
Observations	1,204,200	1,204,200

Notes: Dependent variable is the proportion of repeat matches at school-subject-grade-year level. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

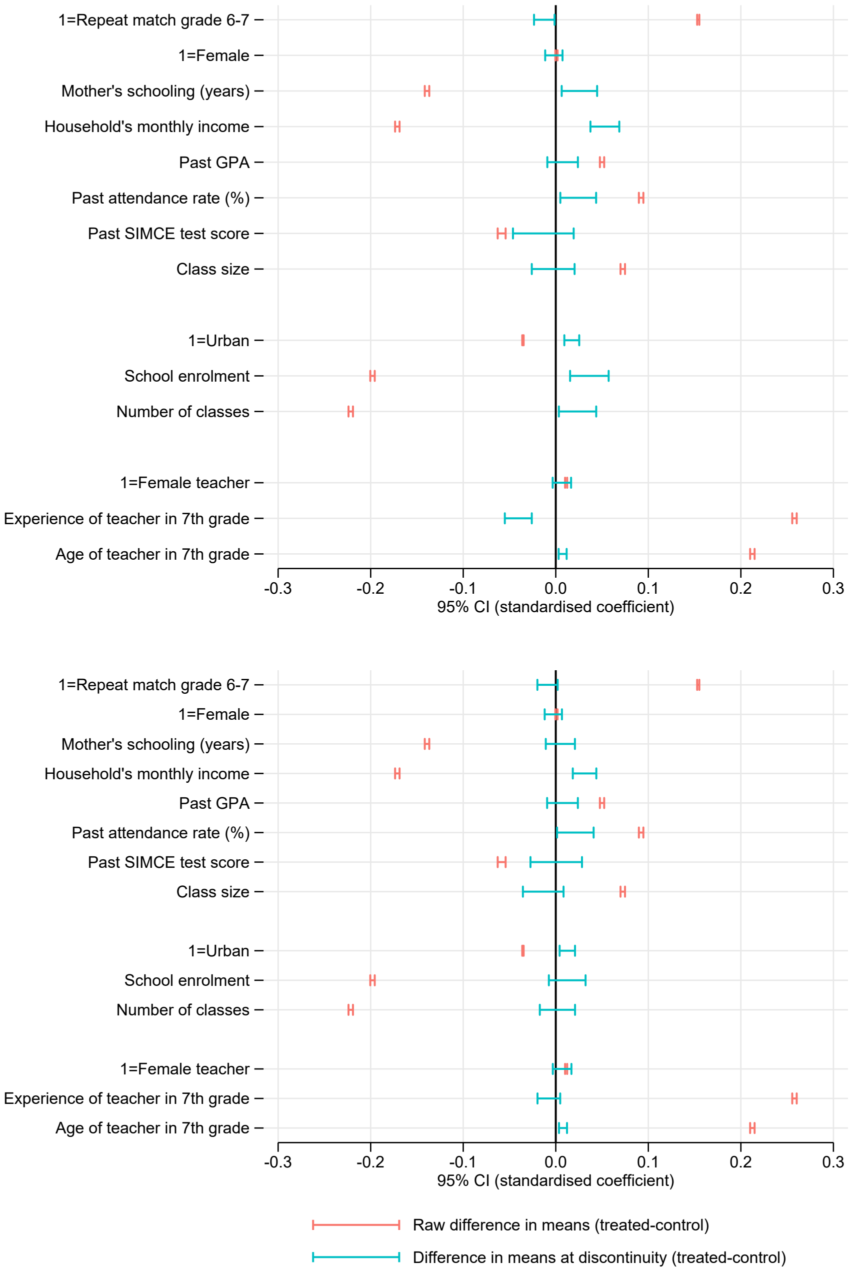
B.2 The exogeneity of the LRA discontinuity

TABLE B.2. Tests of balance of income bands at the discontinuity, controlling by school type

Covariate	RD estimator (1)	Std. err. (2)	Obs. left (3)	Obs. right (4)	Bandwidth (5)
Income level: 1	0.001	0.004	232,241	116,886	1098.770
Income level: 2	0.004	0.004	234,639	117,513	1110.402
Income level: 3	0.002	0.004	243,495	119,672	1152.777
Income level: 4	-0.004	0.003	205,880	110,872	987.444
Income level: 5	-0.000	0.002	269,131	124,478	1267.990
Income level: 6	-0.002	0.002	241,984	119,275	1143.496
Income level: 7	-0.000	0.001	234,098	117,224	1107.541
Income level: 8	-0.001	0.001	206,427	110,942	990.439
Income level: 9	-0.000	0.001	257,206	122,031	1217.552
Income level: 10	0.000	0.001	300,977	128,863	1387.952
Income level: 11	0.000	0.001	293,698	128,187	1357.590
Income level: 12	-0.000	0.000	237,756	118,151	1127.060
Income level: 13	-0.000	0.000	258,257	122,226	1222.284
Income level: 14	-0.000	0.000	313,007	130,989	1436.014
Income level: 15	-0.001	0.001	123,887	84,820	619.084

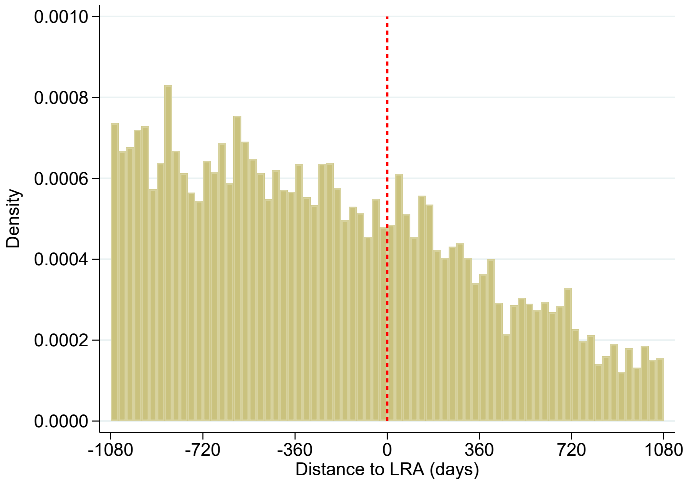
Notes: Table shows the estimated discontinuity in each income band at the LRA controlling for school type. Results based on the empirical strategy that implements a RD following the methodology proposed by Calonico et al. (2014), with a polynomial of order one and weighted by triangular kernel. The number of observations for all the regressions is $2,785,928$. Standard errors calculated using Calonico et al. (2014) and clustered at the student level.

Figure B.2. Balancing tests at the discontinuity



Notes: Figures show 95% confidence intervals on the difference in means between the treated and controls in the overall sample and at the discontinuity in the LRA. All variables are standardised to have zero mean and unit standard deviation to enable comparison. The bottom panel includes as covariates dummies for school type (Public, Private, Voucher). The difference at the discontinuity is estimated using methodology proposed by Calonico et al. (2014), with a polynomial of order one and weighted by triangular kernel. The number of observations for all the regressions is 2, 785, 928. Standard errors calculated using Calonico et al. (2014) and clustered at the student level.

Figure B.3. Density of the running variable



Notes: Running variable is distance, in days, from age on the final day of the grade 7 school year to the day on which the teacher reaches the legal retirement age. Bins have width of 30 days.

B.3 Experience Model

Table B.3. Returns to experience across different experience ranges

	(1)
1–2 years of experience	0.014***
	(0.003)
3–4 years of experience	0.032***
	(0.003)
5–9 years of experience	0.037***
	(0.003)
10–14 years of experience	0.040***
	(0.003)
15–24 years of experience	0.036***
	(0.003)
>25 years of experience	0.024***
	(0.003)
Student FE	Yes
Subject FE	Yes
R-squared	0.793
Observations	2,785,928

Notes: Dependent variable is the student's SIMCE test score in grade 8. The model shows the estimated returns to experience across different experience ranges. The omitted category is teachers with zero experience. The model includes a female teacher dummy. Standard errors are clustered at the student level. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

B.4 Regression discontinuity results by subject

TABLE B.4. Effect of repeated matches on test scores by subject: linear regression discontinuity results

	Spanish	Maths	Natural Sciences	Social Sciences
	(1)	(2)	(3)	(4)
τ_R (First stage)	-0.088***	-0.164***	-0.111***	-0.161***
	(0.003)	(0.004)	(0.004)	(0.004)
τ_y (Reduced form)	-0.009	-0.013	-0.005	-0.014
	(0.010)	(0.012)	(0.012)	(0.011)
τ_{RD}	0.102	0.080	0.049	0.085
	(0.118)	(0.073)	(0.105)	(0.071)
Teacher FE × Year FE	Yes	Yes	Yes	Yes
First-stage R-squared	0.762	0.753	0.760	0.765
First-stage F statistic	654	1,498	780	1,826
Observations	696,482	696,482	696,482	696,482

Notes: Dependent variable is the student's SIMCE test score in grade 8 in each subject. Treatment is the student-subject measure of repeated match R_{is}. All the models therefore control for student characteristics (gender, household income, mother's education, final GPA and lagged attendance rate), class size and school characteristics (public school dummy and rural school indicator). $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

B.5 Information from teachers' survey

Table B.5. Mean comparison test of classroom characteristics, full sample versus estimation sample

	Estimation sample	Sample with teachers' perception	Difference	Std. err.
1=Female	0.50	0.50	-0.005**	(0.002)
Mother's schooling (years)	10.67	11.12	-0.443***	(0.032)
Household's monthly income	378.51	437.00	-58.486***	(5.033)
Average SIMCE test score	-0.06	-0.03	-0.028***	(0.007)
Past GPA	0.08	0.09	-0.017***	(0.004)
Past attendance rate	94.34	94.08	0.257***	(0.041)
Class size	21.88	21.02	0.855***	(0.120)
1=Public	0.53	0.44	0.084***	(0.006)
1=Urban	0.82	0.82	-0.008*	(0.005)
Observations	31,837	9,498		

Notes: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Appendix C

C.1 Data availability

Table C.1. Data available for SIMCE test scores and parental investment

Cohort - High school graduates	4th grade	8th grade	10th grade
2004	-	Yes	-
2005	-	-	Yes
2006	-	-	-
2007	-	-	-
2008	-	Yes	Yes
2009	-	-	-
2010	Yes	-	Yes
2011	-	Yes	-
2012	-	-	Yes
2013	Yes	Yes	-
2014	Yes	-	Yes
2015	Yes	Yes	Yes
2016	Yes	-	Yes
2017	Yes	Yes	Yes
2018	Yes	Yes	Yes
2019	Yes	Yes	Yes

Notes: Cohorts with information on parental investment in blue.

C.2 School starting age and student outcomes

Table C.2. The effect of own school starting age on student outcomes, by birth order

	Test scores in 4th grade	Test scores in 8th grade	Test scores in 10th grade	High school grades	College enrolment
	(1)	(2)	(3)	(4)	(5)
Younger sibling:	0.183***	0.139***	0.128***	0.098***	0.073***
$\mathbf{1}(B_i > 0)$	(0.021)	(0.021)	(0.024)	(0.020)	(0.009)
95% C.I.	[.142 ; .225]	[.097 ; .18]	[.081 ; .176]	[.058 ; .137]	[.056 ; .09]
Effective obs.: Left	20,731	15,489	20,152	27,005	21,202
Effective obs.: Right	20,445	14,751	19,075	25,487	20,225
Optimal Bandwidth	49.58	45.37	47.79	39.45	31.76
Observations	161,177	129,958	161,859	260,393	258,274
Older sibling:	0.130***	0.096***	0.071***	0.033**	0.017*
$\mathbf{1}(B_i > 0)$	(0.048)	(0.023)	(0.021)	(0.015)	(0.010)
95% C.I.	[.036 ; .224]	[.051 ; .142]	[.03 ; .112]	[.004 ; .063]	[-.002 ; .036]
Effective obs.: Left	5,820	10,541	16,206	38,008	38,100
Effective obs.: Right	7,400	12,141	18,632	43,480	44,082
Optimal Bandwidth	42.64	43.83	50.99	57.76	65.97
Observations	54,145	94,954	126,450	260,393	230,397

Notes: Table shows the effects of school starting age on different outcomes. Results based on the empirical strategy that implements a RD following the methodology proposed by Calonico et al. (2014), with a polynomial of order one and weighted by triangular kernel. Standard errors are clustered at the running variable at daily level. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

C.3 Sibling school starting age and student outcomes

Table C.3. The effect of sibling school starting age on student outcomes

	Test scores in 4th grade (1)	Test scores in 8th grade (2)	Test scores in 10th grade (3)	High-school grades (4)	College enrolment (5)
Older-to-younger:	-0.010	-0.041	-0.013	-0.026*	-0.001
$\mathbf{1}(B_j > 0)$	(0.016)	(0.026)	(0.020)	(0.014)	(0.009)
95% C.I.	[-.041 ; .022]	[-.092 ; .01]	[-.051 ; .026]	[-.055 ; .002]	[-.019 ; .016]
Effective obs.: Left	15,609	14,656	25,291	30,745	35,767
Effective obs.: Right	17,171	16,621	28,564	35,025	40,704
Optimal Bandwidth	37.72	44.93	61.96	46.72	54.91
Observations	161,177	129,958	161,859	260,393	258,274
Younger-to-older:	0.039	0.072**	0.071**	0.015*	0.025*
$\mathbf{1}(B_j > 0)$	(0.032)	(0.034)	(0.032)	(0.022)	(0.014)
95% C.I.	[-.024 ; .102]	[.005 ; .139]	[.008 ; .134]	[-.028 ; .058]	[-.002 ; .052]
Effective obs.: Left	9,495	10,241	16,831	38,718	24,903
Effective obs.: Right	7,936	9,391	15,716	36,815	23,762
Optimal Bandwidth	63.25	40.58	50.32	56.45	41.96
Observations	54,145	94,954	126,450	260,393	230,397

Notes: Table shows the effects of sibling school starting age on different outcomes. Results based on the empirical strategy that implements a RD following the methodology proposed by Calonico et al. (2014), with a polynomial of order one and weighted by triangular kernel. Standard errors are clustered at the running variable at daily level. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

C.4 Heterogeneous effects

Table C.4. The effect of sibling school starting age on PSU test scores by student characteristics

	Older-to-younger				Younger-to-older			
	RD (1)	Std. err. (2)	Bandwidth (3)	Obs. (4)	RD (5)	Std. err. (6)	Bandwidth (7)	Obs. (8)
Gender								
Female	0.005	0.016	67.81	50,173	0.077***	0.023	36.63	25,099
Male	-0.011	0.022	57.98	38,815	0.043*	0.024	48.43	31,053
Gender composition								
Same-gender	-0.018	0.018	48.79	37,048	0.039**	0.017	60.35	43,652
Opposite-gender	0.005	0.017	51.99	33,558	0.031	0.020	65.10	40,523
Income								
Below median	-0.001	0.019	45.04	32,248	0.028	0.018	46.12	30,828
Above median	-0.009	0.019	49.54	34,964	0.057**	0.023	53.82	35,720
Father completed HS								
No	-0.024	0.023	45.42	18,197	-0.017	0.020	66.92	25,368
Yes	-0.002	0.015	50.79	51,282	0.038**	0.017	69.81	67,126
Mother completed HS								
No	-0.022	0.024	55.66	21,388	0.028	0.022	58.87	21,593
Yes	0.005	0.013	57.23	59,351	0.038**	0.016	65.99	63,692
Timing								
Grade 1 – 4					0.091***	0.022	34.86	27,001
Grade 5 – 12					0.005	0.025	40.83	22,033
GPA$_{t-1}$								
Low					0.007	0.027	40.24	9,351
Medium					0.014	0.029	64.25	14,651
High					0.097***	0.035	43.67	9,798

Notes: Table shows show the heterogeneous effects for younger (older-to-younger) and older siblings (younger-to-older) by student characteristics. Results are based on linear regressions within the optimal bandwidth computed using the methodology proposed by Calonico et al. (2014) and reported in Table 3.8. Standard errors are clustered at the running variable at daily level. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

C.5 Potential channels: parental investments

Table C.5. Parental investments and sibling school starting age (using raw measures)

	Older-to-younger	Younger-to-older
	(1)	(2)
(a) 4th grade (Parent Responses)		
1=Read to child	0.005	0.168
	(0.081)	(0.112)
1=Parent–child joint reading	-0.039	0.205
	(0.083)	(0.149)
1=Talk about their readings	0.011	0.104
	(0.066)	(0.086)
PSU test scores	-0.075	0.191*
	(0.077)	(0.107)
Observations	14,026	5,809
(b) 8th grade (Student Responses)		
1=Parent congrats for grades	0.042	0.095**
	(0.032)	(0.041)
1=Parent knows grades in school	0.017	-0.023
	(0.038)	(0.055)
1=Parent willing to help	-0.003	0.133***
	(0.028)	(0.049)
PSU test scores	-0.014	0.111**
	(0.031)	(0.053)
Observations	71,769	27,810

Notes: Table shows the effect of sibling school starting age on parental investments in grades 4 and 8 for younger siblings (older-to-younger) – Column (1) – and older siblings (younger-to-older) – Column (2). Results are based on the empirical strategy that implements a RD following the methodology proposed by Calonico et al. (2014), with a polynomial of order one and weighted by triangular kernel. Standard errors are clustered at the running variable at daily level. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.