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

Facundo Albornoz[*]

University of Nottingham

David Contreras[†]

University of Nottingham

Richard Upward[‡]

University of Nottingham

March 1, 2021

[*]facundo.albornoz@nottingham.ac.uk. Address: Room B66a Sir Clive Granger Building, University Park Nottingham, UK. Phone: +44 115 95 15474.

[†]david.contrerasgomez@nottingham.ac.uk. Address: Room B49 Sir Clive Granger Building, University Park Nottingham, UK. Phone: +44 755 146 1260.

[‡]Corresponding author. richard.upward@nottingham.ac.uk. Address: Room B3 Sir Clive Granger Building, University Park Nottingham, UK. Phone: +44 115 95 14735.

Abstract

We use rich, comprehensive, student-teacher data to explore the effectiveness 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. We exploit 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. Repeating the student-teacher match has a robust positive effect on test scores which aggregates up to the class and school-level, and which can also be observed in university entry test scores taken at the end of a student's school career. Repeating the match also has a positive effect on student behaviour and teacher expectations.

JEL classification: I21, I25

Keywords: student-teacher matches, student achievement, looping

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 specialize 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 paper 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 & Dichanz 1994), in Chinese schools at all levels (Liu 1997) as well as in Finland, Japan, Sweden, Israel and Italy (Tourigny, Plante & Raby 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 paper, 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

[1] Hill & 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, Hanushek & Kain (2005) and Rockoff (2004).

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 class or school level. Using the same regression discontinuity methods, we find class-level effects which are very similar to the student-subject-level effects. At the school-level, a discontinuity approach is not possible and so we rely on fixed effects methods which account for possible selection at the school-year and subject-level. Reassuringly, we find school-level estimates to be slightly larger than the equivalent student-subject level estimates. In another exercise, we also show that exposure to looping has positive impacts beyond the year in which the repeat match takes place. We show that students who are exposed to more repeat-matches during their school career perform better on their university entry test taken at the end of their high-school career.

Finally, we explore a plausible potential channel 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 emphasized the positive relationship between school effectiveness and a cooperative 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, Cohen, Guffey & Higgins-D'Alessandro 2013). In addition, recent studies (Bryk, Sebring, Allensworth, Easton & Luppescu 2010, Kraft, Marinell & Shen-Wei Yee 2016, Klugman 2017) have established a positive causal impact of school climate on students achievement on standardized test scores.[3] We find that that in classes with more student-teacher matches, 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 & 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 complement 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 investigates 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

[3]For a comprehensive review on school climate literature, see Thapa, Cohen, Guffey & Higgins-D'Alessandro (2013).

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

evidence on how student-teacher familiarity manifests in better classroom behaviour.

A recent literature emphasises complementarities between teacher and student characteristics (e.g. Aucejo, Coate, Fruehwirth, Kelly & Mozenter 2018, Graham, Ridder, Thiemann & Zamarro 2020). This implies that improving teaching-to-classroom assignments may lead to better student outcomes. Graham, Ridder, Thiemann & Zamarro (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 paper 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 & Nichols (2002), Cistone & Shneyderman (2004), Tucker (2006) and Franz, Thompson, Fuller, Hare, Miller & Walker (2010). Cistone & 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 & 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 & 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 & Dacus (1999) and Black (2000). A teacher cited by Little & 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. However, looping may also have disadvantages. For example, Bogart (2002) notes that some unlucky students may spend two or more years with an ineffective teacher, and teachers may find it more difficult to teach a multi-year rather than a single-year 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 paper is a contribution in that direction.

The remainder of the paper is organized as follows. Section 2 describes our data, and Section 3 explains the econometric framework, including evidence that the discontinuity at the LRA provides a relevant and exogenous instrument. In Section 4 we estimate the short-run effects of repeated matches on individual student-subject outcomes, and in Section 5 we estimate short-run effects at class and school level to see whether the student-subject level effects aggregate up. In Section 6 we estimate longer-run effects over the course of a student's school career. In Section 7 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 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) 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 standardize 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 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 standardized test administered by the Ministry of Education to all students in certain grades, and 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 standardized 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]

[5]We do not use these school grades as an outcome measure because they may reflect teacher biases as well as student performance (Contreras 2019).

[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 individuals students cannot attend.

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.

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

[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.

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 3.

Table 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, 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

[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 will 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.

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

[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 and 9% (4,151 teachers) are observed four times.

[12]For a detailed description of the Chilean school system and education providers, see Santiago, Fiszbein, Jaramillo & Radinger (2017).

Table 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.

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 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.

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. 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 anonymized SIMCE test score.[15]

Panel (c) of Table 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

[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 were coded as dummies 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 was 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.

[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 standardized to have mean zero and unit variance.

Table 2. Characteristics of treatment and control groups

	Treatment group (same teacher in grade 8)	Control group (new teacher in grade 8)	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. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

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). 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.

3 Methods

The first aim of this study is to measure the causal effect of having a repeat match on students' standardized test scores. As shown in Table 2, a simple comparison of repeat matches and new matches may be misleading because repeat matches are not randomly

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

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, Ladd & Vigdor 2005, 2006, Feng 2010, Kalogrides, Loeb & Béteille 2013). Moreover, qualitative research shows that school leaders base their staffing decisions on a combination of teachers' performance (measured by their students' test scores) and teachers' preferences (Cohen-Vogel 2011, Kalogrides et al. 2013, Osborne-Lampkin & 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

[17]Many cross-sectional studies exploit within-student variation to identify effects of teacher characteristics and teaching practises (Dee 2007, Clotfelter, Ladd & Vigdor 2010, Bietenbeck 2014, Bietenbeck, Piopiunik & Wiederhold 2018, Paredes 2014, Lavy 2015, Comi, Argentin, Gui, Origo & Pagani 2017).

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 assigning 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 allow 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, 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). 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:

$$y_{is} = \beta_1 R_{is} + \mu_i + \mu_s + \mu_{jt} + \varepsilon_{is}, \tag{1}$$

where y_{is} is the standardized 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 $t = 2004, 2007, 2009, 2011$. For a particular student-subject-year combination we observe the identity $j = J(i,s,t)$ of the teacher. Each student i is observed in only one school, whereas teachers j may be observed in several schools. 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 include student, subject and teacher-by-year fixed effects.[18]

The remaining source of variation in (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. There are at least three potential sources of selection on match quality.

A first possible source of selection on ε_{is} might be parental choice, if parents choose a particular class within a school on the basis of the quality of the student-teacher match in a particular subject. However, we note that only 2.8% of students change class within the same school between grade 7 and grade 8. Second, and perhaps more seriously, schools may make decisions about which class-teacher matches to keep together in grade 8 on the basis of their performance in grade 7. Finally, selection may also arise if repeat matches have effects which persist over time. For example, suppose that classes which have a repeat match between $t-2$ and $t-1$ are systematically more or less likely

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

to have a repeat match between $t-1$ and t. If repeated matches have any long-lasting effect, then $R_{is,t-1}$ will be part of ε_{is} and will be correlated with $R_{is,t}$.

One solution is to estimate a value-added version of (1) which controls for lagged test scores at the student-subject level (Rivkin et al. 2005, Harris & Sass 2011, Chetty, Friedman & Rockoff 2014). However, the SIMCE test score information for these students is only available in 4th and 8th grade, which means we can only identify the cumulative effect of repeat matches in grades 5 to 8, so this model is not directly comparable to (1). For the same reason, this model does not include teacher fixed effects because each student will have multiple teachers for each subject between 5th and 8th grades. Nevertheless, the value-added model is a useful test of whether the estimates from (1) are subject to within-subject selection bias. The value-added model we estimate is

$$y_{is} = \beta_1 \sum_{g=5}^{8} R_{isg} + \beta_2 y_{is,4} + \mu_i + \mu_s + \varepsilon_{is}, \tag{2}$$

where $\sum_{g=5}^{8} R_{isg}$ is the number of classes in subject s in which student has a repeat match between grades 5 and 8. This model imposes a linear effect on repeat matches which can be relaxed by estimating a separate effect for 1, 2, 3, ... repeat matches. Although the value-added approach controls for student-subject level grade 4 test scores, it remains the case that shocks to student-subject match quality which form part of ε_{is} in (2) may be correlated with the number of repeat matches a student experiences between grades 5 to 8. In addition, (2) does not allow for selection into repeat matches on the basis of teacher characteristics.

We therefore 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 & Lovenheim (2014). Hanushek, Kain & Rivkin (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.[19] 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 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. Figure 1 illustrates that the probability of retiring, using this measure, is strongly and discontinuously related to the LRA. There appears to be very little early retirement. There is a small increase in the probability of retirement in the year before the LRA

[19]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.

is reached, but apart from this, the probability of retirement is only weakly related to age until the LRA is reached, at which point there is a large and sudden increase in the probability of retirement.

In Figure 2 we show how the discontinuity in retirement at the LRA also causes a discontinuity in the probability of having a repeat match. Here, we define the distance to the LRA as the difference between the current age and the LRA, which allows us to combine female and male teachers together. There is a strong positive relationship between teacher age and the probability of a repeat match, but the probability of repeating the student-teacher match drops suddenly at the retirement threshold.

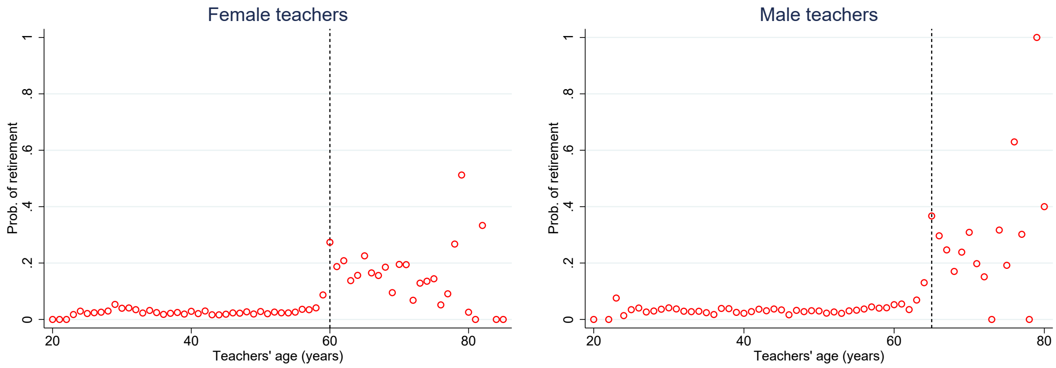
There is some indication in Figures 1 and 2 that retirement is more likely in the year before the LRA is reached, which seems plausible if teachers are unable or unwilling to retire in the middle of a school year. In that case, a teacher who will reach the LRA in the next school year may choose to retire before the school year starts rather than wait until the end of that school year. However, this small discontinuity disappears if we examine distance to the LRA measured in months, shown in Figure 3. Here we see evidence 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$=-1$). This discontinuity is then reflected in an approximately 15 percentage point reduction in the probability of a repeat match, shown in the right hand panel.

Formally, for this discontinuity to identify a first stage we require that:

$$\lim_{D_{is}\downarrow 0} \Pr(R_{is}=1|D_{is}=0) \neq \lim_{D_{is}\uparrow 0} \Pr(R_{is}=1|D_{is}=0)$$

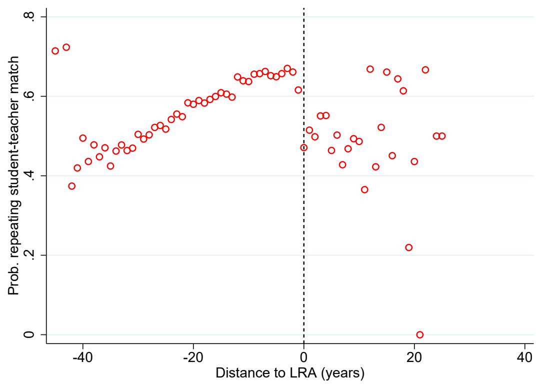
where R_{is} denotes the treatment variable (repeated student-teacher match) for student i for subject s, D_{is} is the running variable, which in this case is distance to the LRA on the last day in the grade 7 school year (end of February). The running variable is

Figure 1. Probability of retirement and teachers' age, age in years



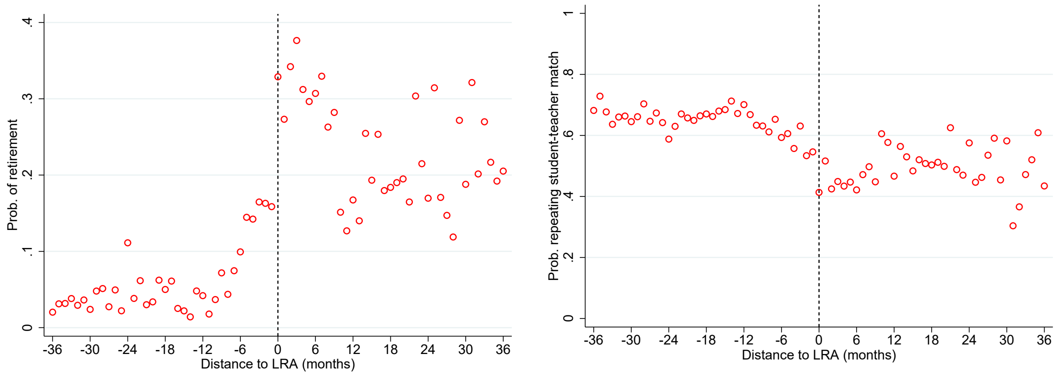
Notes: A teacher is considered retired if she does not appear in the next five consecutive years on the administrative records of Ministry of Education. The dashed line corresponds to the LRA (60 years for women and 65 for men). The teacher's age is grouped by bins of 1 year. The dots are the means of the probability of retirement within each bin.

Figure 2. Probability of repeating the student-teacher match and distance to the LRA, distance in years



Notes: The distance to the LRA is the difference between the current age on the last day in February and the LRA. The distance to LRA is grouped grouped by bins of 1 year. The dashed line at zero therefore corresponds to a teacher who reaches the LRA in the grade 7 school year (from March-February). The dots are the means of the outcome variable (probability of repeating the student-teacher match) within each bin.

Figure 3. 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.

recorded in days. As discussed, the probability of treatment discontinuously decreases at the cutoff point: students whose teachers reach the LRA ($D_{is} = 0$) are less likely to repeat the student-teacher match. We have a fuzzy-RD design, because classes whose teacher reaches the retirement age may still have a repeat match because their teachers choose to delay retirement ("always-takers") and classes whose teacher does not reach the retirement age may not have a repeat match ("never-takers").

Following Imbens & Lemieux (2008) the RD estimator is defined as:

$$\tau_{RD} = \frac{\lim_{D_{is}\downarrow 0} E[y_{is}|D_{is}=0] - \lim_{D_{is}\uparrow 0} E[y_{is}|D_{is}=0]}{lim_{D_{is}\downarrow 0} E[R_{is}|D_{is}=0] - \lim_{D_{is}\uparrow 0} E[R_{is}|D_{is}=0]} = \frac{\tau_y}{\tau_R} \tag{3}$$

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 the interval $[-h, +h]$. We use a triangular kernel function, which gives the maximum weight at $D_{is} = 0$. Finally, 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, Cattaneo & Titiunik (2014) by selecting the parameter h that minimizes 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). This procedure removes approximation error of the local polynomial estimator ($\hat{B}_n$) from the RD point estimator ($\hat{\tau}^{RD}$), leading the bias corrected point estimator to be re-centred at τ^{RD}. By using this correction is possible to compute an alternative asymptotic approximation to the bias-corrected RD estimator, resulting in an asymptotic variance V_n^{bc}. This leads to the robust bias-corrected confidence intervals:

$$I^{rbc} = [(\hat{\tau}^{RD} - \hat{B}_n) + \Phi^{-1}_{1-\frac{\alpha}{2}} \sqrt{V_n^{bc}}] \tag{4}$$

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 A1 in Appendix A 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 A2 shows that density of the running variable shows no sign of manipulation at the cutoff. 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. In this setting, the exclusion restriction appears quite plausible, since we are comparing outcomes between students whose teachers' age varied slightly *in the previous school year*. However, there is one important caveat. 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. 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 (1).

4 The effects of repeated matches at the student-subject level

In this section we consider the effect of repeated student-teacher matches on test scores measured at the student-subject level, as in Equation (1). In the next section we consider effects of repeated matches at the 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.

Table 3 presents estimates of versions of Equation (1) with the inclusion of different fixed-effects. Across all four specifications, the results show a positive and significant effect of repeating the student-teacher match on student's SIMCE test scores. The raw

Table 3. Effect of repeat student-teacher match on test scores: fixed-effect estimates

	(1)	*(2)*	*(3)*	*(4)*
Repeat match	0.003**	0.026***	0.017***	0.019***
$R_{is} = 1$	(0.002)	(0.001)	(0.001)	(0.002)
Subject FE	Yes	Yes	Yes	Yes
Student FE		Yes	Yes	Yes
Teacher FE			Yes	
Teacher FE × Year FE				Yes
R-squared	0.000	0.793	0.808	0.812
Observations	2,785,928	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}. Standard errors are clustered at the student level. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

effect in Column (1) is small, but recall from Table 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 somewhat. 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. Our estimates are very similar to those reported by Hill & Jones (2018) for younger students' maths scores in North Carolina elementary schools (grades 3–5) using a similar specification, but which also includes lagged test scores as a control variable.[20] The outcome measure used by Hill & Jones is a maths score which was reported by the teacher themselves, rather than an anonymized national test score as in our case. This suggests that the use of an anonymized test score, as in our case, is not crucial for finding positive effects from repeated matches.

The fixed-effects regression (1) corrects for potential bias arising from selection of students and teachers, exploiting the within-student and within-teacher-year varia-

[20] Hill & Jones (Table 2) 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 & Jones report that only three percent of students experience a repeat match in their data.

tion to identify the effect. Nevertheless, as explained in Section 3, this strategy does not control for selection bias due to match quality at the student-subject (equivalently student-teacher) level. We therefore also report estimates of the value-added model (2) which controls for subject-specific student achievement.

In panel (a) of Table 4 we show that an additional repeat match increases grade 8 SIMCE test scores by 0.016σ, a similar effect size to that reported in Table 3. In panel (b) we explore the linearity assumption by separating the effect into 1, 2 or ≥ 3 repeat matches. Reassuringly, the effect is monotonically increasing in the number of repeat matches and, in fact, a linear relationship is a close approximation. In columns (2), (3) and (4) we present estimates separately by subject, at the expense of the inclusion of student fixed effects (recall that each student is observed once in each subject.[21] In each case, the number of repeat matches has a positive and significant effect on test scores in grade. The estimated effect size from column (1) is very similar to that estimated by the fixed-effects model in Table 3.

The question remains whether the positive association between repeated matches and test scores is driven by unobserved match-level characteristics. We therefore now consider our regression discontinuity design, illustrated in Figure 4, 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 A (Figure A1) 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

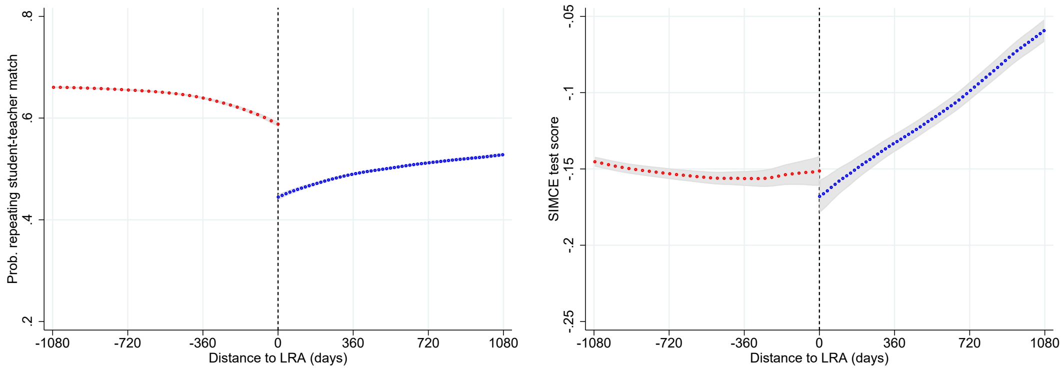
[21] We only have lagged test score information for 3 rather than all four subjects, and for only two years: 2009 and 2011.

Table 4. Effects of repeated matches on SIMCE test scores: Value-added approach

	All	*Spanish*	*Maths*	*Natural Sciences*
	(1)	*(2)*	*(3)*	*(4)*
(a) Linear Model				
Number of matches	0.016^{***}	0.004^{*}	0.011^{***}	0.008^{***}
grade 5–8	(0.001)	(0.002)	(0.002)	(0.002)
Lagged SIMCE score	0.283^{***}	0.655^{***}	0.650^{***}	0.632^{***}
	(0.002)	(0.002)	(0.002)	(0.002)
R-squared	0.829	0.580	0.642	0.601
(b) Non-linear Model				
$R_i = 1$	0.018^{***}	0.008	0.016^{***}	0.019^{***}
	(0.003)	(0.005)	(0.005)	(0.005)
$R_i = 2$	0.031^{***}	0.004	0.025^{***}	0.021^{***}
	(0.003)	(0.006)	(0.005)	(0.006)
$R_i \geq 3$	0.052^{***}	0.015^{**}	0.035^{***}	0.028^{***}
	(0.003)	(0.007)	(0.007)	(0.007)
Lagged SIMCE score	0.283^{***}	0.655^{***}	0.650^{***}	0.632^{***}
	(0.002)	(0.002)	(0.002)	(0.002)
Student FE	Yes			
Subject FE	Yes			
School FE × Year FE		Yes	Yes	Yes
R-squared	0.829	0.580	0.642	0.601
Observations	759,597	253,199	253,199	253,199

Notes: Dependent variable is the student's SIMCE test score in grade 8. Panel A: Treatment is the number of matches between grade 5 and grade 8 in each subject. Panel B: Treatment is a set of dummies that represents the number of repeat-matches between grade 5 and grade 8. Columns (2)-(4) control for student characteristics (gender, household income, mother's education and attendance rate) and class size. Standard errors are clustered at the student level in Column 1. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Figure 4. 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.

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, 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.[22] 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 A1. 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.

[22]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.

Figure 4 implies a causal effect of repeat matches which is substantially larger than the fixed-effect estimates in Table 3, because the ratio of τ_y and τ_R is approximately 0.2σ. In Column (1) of Table 5 we report a non-parametric RD estimate of 0.158σ which corresponds exactly to Figure 4. 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 5 we we adopt a linear functional form for the distance to the LRA, which has a number of advantages at this point. 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 column (1) and (2) can be accounted for by the loss of experience which is associated with getting a new teacher in grade 8.[23]

A disadvantage of the simple linear model reported in column (2) is that Figure 4 suggests that the relationship between looping and age is somewhat non-linear in the

[23]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 & Sass (2011) our return to experience model in Appendix B predicts that losing a teacher at the LRA with 25 years experience (the sample mean) and replacing them with a new teacher causes a loss in student test scores of 0.024σ.

Table 5. 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 $\times$ Year FE		Yes	Yes
First-stage R-squared		0.873	0.873
First-stage F statistic		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 Equation (4). Standard errors are clustered at the student-level. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

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-effects. The quadratic model yields an estimate of 0.124σ, with a slightly larger standard error than the linear model. All our RD estimates are larger than the fixed-effects and value-added estimates, even after allowing for the large experience effect. 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.

A natural question is whether the positive effects of repeated matches occur for every subject. In Table 6 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 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.

Table 6. 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 $\times$ 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 attendance rate), class size and school characteristics (public school dummy and rural school indicator). $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

5 The effect of repeated matches on classes and schools

5.1 Class-level effects

The comparison we made in Table 5 was between individual student-teacher matches that repeat and those that do not. Repeat matches may also have effects on students who do not themselves have the same teacher, but who move into a class in which other students do have the same teacher. If repeat matches allow teachers to save time, there might have benefits on all students in the class. 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.

To test whether repeat matches have benefits which spillover within the class, we collapse the data to class-subject level and create a class-subject level measure of repeated matches, R_{cs} which equals one if the majority of students in class c in subject s in grade 8 have the same teacher as in grade 7.[24] Results from this exercise are reported in Table 7. The results in this table can be compared to those in Table 5. The discontinuity is now based on the distance to the LRA for the majority of students in the class, rather than individual students, but the first stage is still extremely strong. However, the greatly reduced sample size — we now have approximately 60,000 classes rather than several million individual matches — leads to less precise reduced form estimates. The resulting RD estimate is of a similar magnitude to that in Table 5, but with larger standard errors.

[24]In a very small number of cases the number of students who repeat the match is equal to the number who do not repeat; we drop these cases. We also drop singleton observations i.e. observations where there is no variation in teacher at the class-subject level.

Table 7. Effect of repeat matches on test scores at class-subject level

	Non-parametric *(1)*	*Linear with fixed effects* *(2)*	*Quadratic with fixed effects* *(3)*
τ_R (First stage)	-0.172^{***}	-0.182^{***}	-0.140^{***}
	(0.027)	(0.021)	(0.025)
τ_y (Reduced form)	-0.024	-0.013	-0.020^{**}
	(0.028)	(0.008)	(0.009)
τ_{RD}	0.139	0.069	0.144^{**}
	(0.155)	(0.044)	(0.071)
Class FE		Yes	Yes
Subject FE		Yes	Yes
Teacher FE		Yes	Yes
First-stage R-squared		0.772	0.773
First-stage F statistic		73	32
95% C.I.	[-.166 ; .444]		
Effective observations: Left	7,421		
Effective observations: Right	2,850		
Optimal Bandwidth	1375.224		
Observations	60,692	60,692	60,692

Notes: Dependent variable is the average SIMCE test score at class-subject level in grade 8. Treatment is the class-subject-level measure of repeated match R_{cs} which takes the value 1 if the majority of students in a class-subject have the same teacher as in grade 7. Column (1) presents results based on Calonico et al. (2014) with a polynomial of order one and weighted by a triangular kernel. Columns (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 Equation (4). Standard errors are clustered at the class-level. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

5.2 School-level effects

We also evaluate whether the prevalence of repeat matches has an effect on school-level outcomes. The allocation of teachers to classes is a joint decision at the school-level, since the decision to repeat one teacher-class match has some implication for all other allocations within that school. We are therefore interested in estimating whether the prevalence of repeat-matches at the school-level has an impact on student test scores.

In Appendix C we provide some evidence on the extent to which repeat matches can be regarded as a policy implemented at the school-level. We find that school fixed-effects can only explain a small fraction of the variation in the prevalence of repeat matches. Instead, the decision to repeat a match appears to be one taken at a lower level i.e. for a particular group of students, a particular teacher, or a particular combination of the two.

At the school-level our RD strategy is not applicable because, although schools will differ in the number of teachers who reach the LRA in each year, there is no sharp discontinuity to exploit, and the number of teachers who reach the LRA depends on the age distribution of teachers in a school, which seems unlikely to be exogenous with respect to student outcomes. We therefore use an aggregated version of Equation (1) which relates the school-subject-year level averages of SIMCE test scores to the proportion of classes in that school-subject-year which are repeat matches. This model allows us to control for school, subject and year fixed effects. We estimate the following equation:

$$\bar{y}_{kst} = \beta_1 \bar{R}_{kst} + \beta_2 S_{ks} + \mu_{kt} + \mu_s + \varepsilon_{ks} \quad (5)$$

where $\bar{y}_{kst}$ and $\bar{R}_{kst}$ are the school-subject-year level averages of y_{is} and R_{is} in Equation (1); μ_{kt} is a school-by-year fixed effect; μ_s is a subject fixed effect; S_{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 (5) has time variation.

Equation (5) 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 attract better students. 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.[25]

Results from estimating (5) are reported in Table 8. Column (1) shows an effect at the school-level which is positive and lies in between the student-subject level estimates reported in Table 3, and the RD estimates reported in Table 5. In columns (2)–(5) we report the effect separately by subject, which suggests a very consistent and positive effect for each subject. We cannot easily compare these to the subject-level estimates reported in Table 6, because the school-subject-level estimates use a different identification strategy, but nevertheless all these results point to a positive effect of repeat matches for all subjects. The fact that the aggregate school-level effect is slightly smaller than the school-subject-level effects is suggestive that there may be negative spillover effects, but these are not large.

6 Effects of repeated matches on a student's school career

All our evidence so far has focused on a short-term effect: the effect on test scores at the end of the year in which the repeated match occurs. Do these effects accumulate over

[25]Table D1 in Appendix D shows a comparison of mean differences of the treatment and the control group within school-by-year. The results suggest that the inclusion of school-by-year fixed effects removes a substantial part of the selection on the basis of student and teacher characteristics.

Table 8. Effect of repeat student-teacher match on test scores at the school-subject level: fixed-effect estimates.

	(1) *All subjects*	*(2)* *Spanish*	*(3)* *Maths*	*(4)* *Natural Sciences*	*(5)* *Social Sciences*
Proportion of repeat matches	0.032^{***}	0.059^{***}	0.060^{***}	0.061^{***}	0.050^{***}
$\bar{R}_{kst} = 1$	(0.003)	(0.009)	(0.008)	(0.009)	(0.008)
School FE $\times$ Year FE	Yes				
Subject FE	Yes				
School FE		Yes	Yes	Yes	Yes
Year FE		Yes	Yes	Yes	Yes
R-squared	0.911	0.772	0.828	0.806	0.815
Observations	82,524	20,631	20,631	20,631	20,631

Notes: Dependent variable is the average SIMCE test score at school-subject level in grade 8. Treatment is the school-subject level proportion of students who have a repeat match in grade 8. Regressions include controls for average teacher experience and proportion of female teachers within the school. Standard errors are clustered at the school level. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

a student's school career? Although we do not observe anonymized test scores in every year, we do observe all student-teacher pairs, so we can construct a measure of how many repeated matches a student is exposed to during their entire middle- and high-school career, and examine whether it has an impact on the University Selection Test (*Prueba de Selección Universitaria* or PSU), which is taken at the end of 12th grade. The PSU consists of two mandatory exams: maths and Spanish, and two optional exams, social sciences and natural sciences. The Chilean Department of Evaluation and Educational Testing Service (*Departamento de Medición, Registro y Evaluación*) has a complete record of PSU test scores. In addition, the dataset contains information about student characteristics, such as gender, family income, household size and working household members.

We focus on five student cohorts of first-time PSU test takers between the years 2014 and 2018. We track these students and their teachers in Spanish and maths, between 5th grade (students aged 10–11) and 12th grade (students aged 17–18). As well as shedding light on possible longer-term effects, our model also allows us to implement a "value-added" specification similar to (2), because, for each student, we have a baseline SIMCE test score recorded in 4th grade, as well as information on student-teacher matches in 4th grade.

The estimation sample consists of students with valid PSU scores for maths and Spanish, a complete history student-teacher match history from 5th grade to 12th grade, and SIMCE test scores in 4th grade. The final sample comprises 551,320 students, covering 52% of the first-time test takers between the years 2014 and 2018. Table E1 in Appendix E reports the data availability by admission process.[26] Table E3 reports descriptive statistics for this long-term treatment indicator. We find that, on average, students have repeat matches in slightly more than one-third of their grades between grades 5 and 12. There are no substantial differences by subjects or educational level.

To identify the effect, we again exploit the fact that we have observations on different subjects for the same student, so we can include student fixed-effects. The inclusion of student fixed-effects controls for student-invariant characteristics related to academic outcomes. In addition, to control for subject-specific ability, we include the student's past SIMCE test score in 4th grade. This measure of students' ability is unaffected by the student-teacher matches which occur between the 5th and 12th grades. Our estimating equation is then:

$$PSU_{is} = \beta_1 \sum_{g=5}^{12} R_{isg} + \beta_2 y_{is,4} + \mu_i + \mu_s + \varepsilon_{is} \tag{6}$$

where PSU_{is} is the standardized test score for the student i in subject s, $y_{is,4}$ is the student's SIMCE test score in subject s in 4th grade, μ_i is a student fixed-effect and μ_s is a subject fixed-effect and the treatment is the number of grades in which student i remains with the same teacher in subject s between 5th and 12th grade.

The results of estimating Equation (6) are shown in Table 9. The estimated effect

[26] To ensure that this subsample is representative, we repeat the model shown in Column (5) of Table 3 for those students who were in 8th grade in 2009 and 2011 and who take the PSU test in admission years 2014 and 2016. Table E2 in Appendix E shows an effect size of 0.024σ, very close to the estimate for the full sample.

size in column (1) is about half the size of the effect from the equivalent value-added model shown in Table 4, which is the effect between grades 5 and 8, possibly some indication that repeated matches may be less effective for older children. However, the effect size for maths shown in column (3) is actually slight larger than the equivalent grade 8 effect. In panel (b) we allow the effect to be non-linear, which again provides monotonically increasing effects which are quite well approximated by a linear model.

7 Classroom behaviour and teacher expectations

Our results consistently show that repeating the student-teacher match results in a positive effect on student test scores both in the short- and long-run. We find these effects at various different levels of aggregation. Further independent evidence of the effectiveness of repeat matches can also be found in 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.

The survey data contains information on teachers' perception of class behaviour and teachers' expectations. 41% of the classes in data have this information for each subject. Table F1 (Appendix F) 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.

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

Table 9. Effects of repeated matches on PSU test scores: Value-added approach

	All (1)	*Spanish* (2)	*Maths* (3)
(a) Linear Model			
Number of matches	0.007***	0.006***	0.014***
grade 5–12	(0.001)	(0.001)	(0.001)
Lagged SIMCE score	0.267***	0.529***	0.417***
	(0.002)	(0.001)	(0.001)
R-squared	0.859	0.595	0.581
(b) Non-linear Model			
$R_i = 1$	0.004	0.002	0.006
	(0.006)	(0.005)	(0.005)
$R_i = 2$	0.010*	0.006	0.007
	(0.006)	(0.005)	(0.005)
$R_i = 3$	0.018***	0.011**	0.024***
	(0.006)	(0.005)	(0.005)
$R_i = 4$	0.024***	0.015***	0.038***
	(0.006)	(0.005)	(0.006)
$R_i = 5$	0.033***	0.026***	0.061***
	(0.006)	(0.006)	(0.006)
$R_i \geq 6$	0.041***	0.036***	0.088***
	(0.008)	(0.007)	(0.007)
Lagged SIMCE score	0.267***	0.529***	0.417***
	(0.002)	(0.001)	(0.001)
R-squared	0.859	0.595	0.581
Student FE	Yes		
Subject FE	Yes		
School FE × Year FE		Yes	Yes
Observations	1,102,640	551,320	551,320

Notes: Dependent variable is the student's PSU test score. Panel A: Treatment is the number of repeat matches between grade 5 and grade 12 in each subject in which a student has a repeat match. Panel B: Treatment is a set of dummies that represents the number of repeat-matches between grade 5 and grade 12. Column (2) and Column (3) control for gender of the student, household income, household number and number of working household members. Standard errors are clustered at the student level in Column (1). $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

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. Then, in order to investigate whether these outcomes are related with the student-teacher match, the following model is estimated:

$$P_{cs} = \beta_1 \bar{R}_{cs} + \beta_2 y_{cs} + \beta_3 T_{cs} + \mu_c + \mu_s + \varepsilon_{cs} \tag{7}$$

where P_{cs} is our measure of teacher perception (behaviour, expectations) for class c subject s, $\bar{R}_{cs}$ is proportion of the class c that repeat the match in the subject s, y_{cs} is the average performance of the class in the SIMCE test in the subject, μ_c is a class fixed effect, μ_s is a subject fixed effect, and T_{cs} are teacher characteristics: gender and experience. Fixed effects at class level are included to capture all the subject-invariant characteristics (observable and unobservable) of the class. The variable $\bar{R}_{cs}$ ranges between 0 and 1, therefore, the coefficient must be interpreted as the effect of the entire class repeating the match on the outcome of interest.

The results are displayed in Table 10. The regression results indicate that repeat matches have a positive effect on the teacher's perception of classroom behaviour and teacher expectations. In particular, teachers are 3.8 percentage points less likely to have behavioural problems at the beginning of the class and 4 percentage points less likely to experience disruptive student behaviour. There are smaller but still significant effects on teacher expectations: teachers are 2 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

Table 10. Effects of repeat matches on student behaviour and teacher expectations

	Problems to start the class	Classroom disruption	High teacher expectations
$\%\text{Match}_{cs}(\bar{R}_{cs})$	-0.038^{***}	-0.040^{***}	0.016^{**}
	(0.007)	(0.008)	(0.007)
Class FE	Yes	Yes	Yes
Subject FE	Yes	Yes	Yes
R-squared	0.419	0.440	0.567
Observations	37,992	37,992	37,992

Notes: All the regressions include average score of the class in the SIMCE test in the subject and teacher characteristics (gender and years of experience). Standard errors are clustered at the class level. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

improves. Of course, we cannot 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.

8 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 paper, 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. Our final piece of evidence suggests that the effects continue over time, and that university test scores increase with the number of repeated matches over a student's school career. 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.[27]

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.

[27] 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.

References

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

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

Bietenbeck, J., Piopiunik, M. & 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.

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

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

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. & Titiunik, R. (2014), 'Robust nonparametric confidence intervals for regression-discontinuity designs', *Econometrica* **82**(6), 2295–2326.

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

Cistone, P. & 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. & 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. & 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. & 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. & Pagani, L. (2017), 'Is it the way they use it? Teachers, ICT and student achievement', *Economics of Education Review* **56**, 24–39.

Contreras, D. (2019), Gender differences in grading: teacher bias or student behaviour? Mimeo, University of Nottingham.

Correia, S. (2016), Linear models with high-dimensional fixed effects: An efficient and feasible estimator, Technical report. http://scorreia.com/research/hdfe.pdf.

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

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

Fitzpatrick, M. D. & 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. & Walker, J. (2010), 'Evaluating mathematics achievement of middle school students in a looping environment', *School Science and Mathematics* **110**(6), 298–308.

Fryer, R. G. J. (2018), 'The "pupil" factory: specialization and the production of human capital in schools', *American Economic Review* **108**(3), 616–56.

Graham, B. S., Ridder, G., Thiemann, P. M. & Zamarro, G. (2020), Teacher-to-classroom assignment and student achievement. NBER Working Paper No. 27543.

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

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

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

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

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

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

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. & 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.

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.

Little, T. S. & 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.

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

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

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.

Rivkin, S. G., Hanushek, E. A. & 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.

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

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

Tourigny, R., Plante, I. & 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.

Zahorik, J. A. & Dichanz, H. (1994), ‘Teaching for understanding in German schools.’, *Educational Leadership* **51**(5), 75–77.

Appendix A The exogeneity of the LRA discontinuity

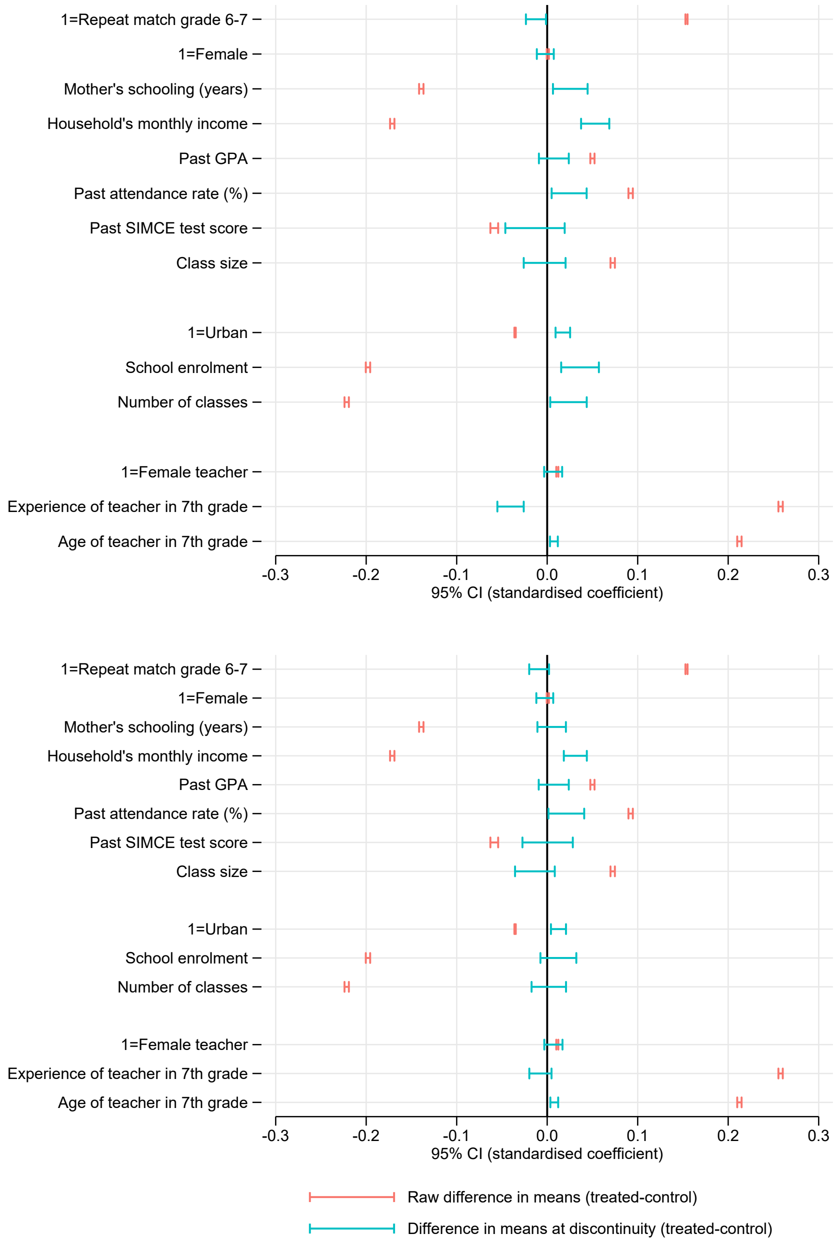


Figure A1. Balancing tests at the discontinuity

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 Equation (4) and clustered at the student level.

Table A1. 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 Equation (4) and clustered at the student level.

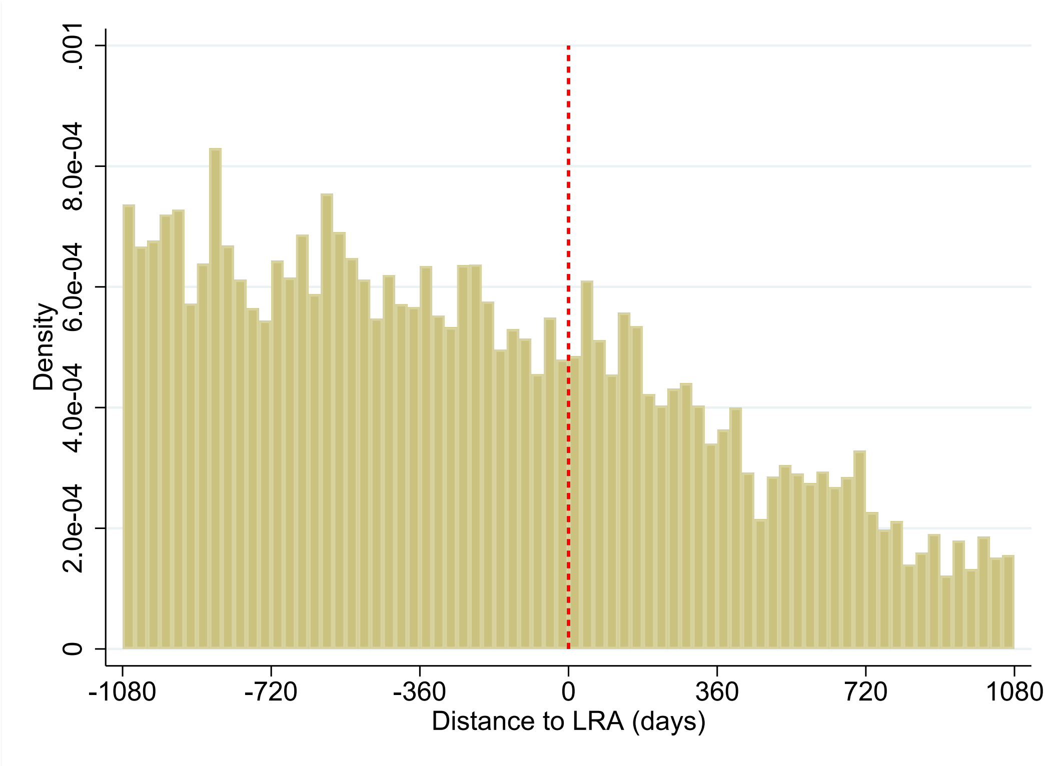


Figure A2. Density of the running variable

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.

Appendix B Experience Model

Table B1. 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. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Appendix C 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 12,111,666 students, 10,036 schools and 455,437 classes.

We aggregate this data to the school level by calculating the proportion of repeat matches in each school across subjects and grades, $\bar{R}_{kt}$. At this level, the sample contains 127,570 observations spanning 10,036 schools observed 12.71 times on average. A variance decomposition exercise reveals that the variation in looping within schools (over time) is almost exactly equal to the variation in average looping behaviour between schools.

We then collapse the raw 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:

$$\bar{R}_{ksgt} = \mu_k + \mu_s + \mu_g + \mu_t \tag{8}$$

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 C1 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 16%, from which we conclude that the prevalence of looping is only weakly associated with school-level decisions.

Table C1. Contribution of school fixed effects to the proportion of repeat matches

	(1)	(2)
1=Maths	0.022^{***}	0.022^{***}
1=Natural Sciences	0.020^{***}	0.020^{***}
1=Social Sciences	0.013^{***}	0.013^{***}
1=Grade 7	-0.144^{***}	-0.071^{***}
1=Grade 8	-0.005^{***}	0.068^{***}
School FE		Yes
Year FE	Yes	Yes
Adjusted R-squared	0.025	0.164
Observations	1,246,936	1,246,936

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

Appendix D Selection within school

Table D1. Characteristics of treatment and control groups within school-by-year

	Treatment group (same teacher in grade 8)	Control group (new teacher in grade 8)	Difference	Std. err.
(a) Student characteristics				
1=Female	0.50	0.51	-0.001^{**}	(0.001)
Mother's schooling (years)	10.94	10.98	-0.036^{***}	(0.004)
Household's monthly income	374.54	378.08	-3.543^{***}	(0.388)
Past GPA	0.12	0.05	0.067^{***}	(0.001)
Past attendance rate (%)	94.58	94.15	0.429^{***}	(0.007)
Past SIMCE test score	0.18	0.17	0.011^{***}	(0.002)
Class size	26.77	26.57	0.201^{***}	(0.004)
(b) Teacher characteristics				
1=Female teacher	0.69	0.68	0.005^{***}	(0.001)
Experience in 8th grade	20.06	16.75	3.306^{***}	(0.012)
Observations	1,618,387	1,167,541		

Notes: The past SIMCE test score is not available for the full sample, 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, within school-by-year. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Appendix E Long-term effects

Table E1. Data availability by admission process

Admission year	*Full sample*	*Estimation sample*
2014	200,631	104,184
2015	203,935	107,240
2016	208,422	110,918
2017	204,112	115,210
2018	207,521	113,768
Total observations	1,057,621	551,320

Notes: The full sample comprises all first-time test takers by each admission process. The estimation sample comprises all first-time test takers by each admission process with valid PSU scores in Verbal and maths, complete student-teacher match history between grade 4 and grade 12, and SIMCE test scores in grade 4.

Table E2. Effect of repeating the student-teacher match on test scores: PSU takers

	(1)
Repeat match	0.024***
$R_{is} = 1$	(0.004)
Student FE	Yes
Subject FE	Yes
Teacher FE × Year FE	Yes
R-squared	0.803
Observations	611,388

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

Table E3. Proportion of grades 5–12 which have a repeat match

	Mean	*Standard deviation*	*Observations*
Estimation sample	0.36	0.18	1,102,640
By subject:			
Spanish	0.36	0.18	551,320
Mathematics	0.37	0.18	551,320
By educational level:			
Second cycle of primary school (grade 5–8)	0.37	0.25	1,102,640
High school (grade 9–12)	0.36	0.26	1,102,640

Appendix F Information from teachers' survey

Table F1. 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: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.