IDB WORKING PAPER SERIES Nº IDB-WP-1075

The Effects of the Chilean School Accountability System on Teacher Turnover

Gregory Elacqua
Diana Hincapié
Matías Martínez

Inter-American Development Bank
Education Division

November 2019

The Effects of the Chilean School Accountability System on Teacher Turnover

Gregory Elacqua*
Diana Hincapié*
Matías Martínez**

* Inter-American Development Bank
** Northwestern University

Inter-American Development Bank
Education Division

November 2019

Cataloging-in-Publication data provided by the
Inter-American Development Bank
Felipe Herrera Library

Elacqua, Gregory M., 1972-
The effects of the Chilean school accountability system on teacher turnover / Gregory Elacqua, Diana Hincapié, Matías Martínez.
p. cm. — (IDB Working Paper Series; 1075)
Includes bibliographic references.
1. Educational accountability-Chile. 2. Teacher mobility-Chile. 3. Teacher turnover-Chile. I. Hincapié, Diana. II. Martínez, Matías. III. Inter-American Development Bank. Education Division. IV. Title. V. Series.
IDB-WP-1075

Contact information: SCL-EDU@iadb.org

The effects of the Chilean school accountability system on teacher turnover

Gregory Elacqua, Diana Hincapié, Matías Martínez*

November 2019

Abstract

This paper estimates the effects of school accountability on year-to-year teacher mobility in Chile. An accountability program was introduced between 2012 and 2015, which established sanctions for persistently low-performing schools, including the threat of closure if they failed to improve their academic outcomes after four years. Since the low-performance ranking was based on the school's relative position on a set of variables and their corresponding thresholds, we use a Multivariate Regression Discontinuity Design to evaluate the impact of the policy on teacher mobility. Our results indicate that teachers are more likely to leave schools that are labeled as low performing. This effect appears to be relevant only when teachers can move to other schools, as we did not find any effect on the likelihood of teachers leaving the school system. The evidence suggests that the effect on mobility is more pronounced for teachers with less working experience, who teach in two or more schools, were hired with temporary contracts, and achieved lower scores on their college admission tests. Even though mobility appears to have increased among less effective teachers, schools are not hiring new teachers to replace them.

Keywords: School accountability; Teacher mobility; Multivariate regression discontinuity
JEL: I28; J18; I24; J63

* Gregory Elacqua and Diana Hincapié: Inter-American Development Bank; Matias Martinez: Northwestern University.
We are grateful to seminar participants at the Education Seminar of the Inter-American Development Bank and the Latin American and Caribbean Economic Association (LACEA) for helpful discussions and comments. This study uses databases from the Education Quality Agency of Chile. The authors thank the Agency forgiving us access to these data. The findings expressed in this paper are those of the authors and do not necessarily represent the views of the Inter-American Development Bank or Northwestern University.

1. Introduction

In 2008, the Chilean government instituted the Preferential School Subsidy Law (SEP law for its acronym in Spanish: *Subvencion Escolar Preferencial*) that introduced a weighted voucher to provide different levels of funding for students to access the school of their choice. Prior to this reform, Chile had a universal per-pupil flat voucher that paid schools a flat fee for each student that attended school. The per pupil voucher included weights for the location of the school (cost of living and whether the school is located in a rural area) but did not adjust for student socioeconomic characteristics. With the introduction of SEP, the voucher increased its value by 50% for students from the lowest 40% of the income distribution, recognizing the higher costs of educating disadvantaged students.

The introduction of SEP spurred a cottage industry of research to identify the effects of the policy on different outcomes, which has fueled an ongoing debate about the merits of the reform. Most of the research has focused on the effect of the reform on student achievement, consistently finding a positive effect on learning and the narrowing of the socioeconomic achievement gap (Murnane, Waldman, Willett, Bos, & Vegas, 2017; Feigenberg, Rivkin, & Yan, 2017), though there is still debate on the magnitude of the effects.

Some studies have attempted to determine if SEP had an impact on expanding schooling options and whether more disadvantaged families are choosing higher performing or more advantaged schools (Navarro-Palau, 2017; Neilson, 2013; Aguirre, 2017) while others have concentrated on the impact of SEP on student segregation (Gazmuri, 2015). However, most of the empirical literature on SEP has focused on the overall impact of the reform on different outcomes while overlooking how some of the key components of the law may influence the results. Some of these components, for example, increased funding for disadvantaged students, and required schools to develop improvement plans and to provide authorities with a detailed report of their spending. Additionally, SEP introduced a new system of school accountability. For the first time, the policy established sanctions for schools that persistently demonstrated low academic performance.

In this paper, we focus our analysis on the effects of the accountability component of SEP on teacher mobility. The main goal of this component was to motivate changes in the school management and teaching practices that could boost efficiency and improve student learning (Elacqua, Santos, Urbina, & Martínez, 2011). One of the underlying assumptions was that principals and teachers would react to the threat of sanctions by increasing their effort, changing working conditions to make their schools more attractive for highly qualified teachers, and/or dismissing ineffective teachers (Ladd & Zelli, 2002; Elacqua, Martínez, Santos, & Urbina, 2016).

However, the policy was not positively perceived by the education community and faced major resistance.[1] The opposition to the use of standardized testing to hold schools and teachers accountable was rooted in the view that many elements influencing student achievement were out of the teachers' control and that establishing threats and standardized evaluations could diminish their intrinsic motivation to teach (Ryan & Deci, 2000). At the same time, accountability systems

[1] One year after its implementation, over 150 academics and researchers wrote an open letter requesting the Ministry of Education to limit the use of standardized tests to establish consequences on school practices.

focused on identifying low-performing schools are usually accompanied by an increase in the administrative workload, reduced autonomy in teaching decisions, and the potential shift in focus from student learning to student testing (Gjefsen & Gunnes, 2016). Additionally, working in a school that has been publicly labeled as low-performing could be considered a stigma for teachers, and they may try to avoid teaching in these schools (Ravitch, 2010; Rouse, Hannaway, Goldhaber, & Figlio, 2013).

These factors could cause an increase in teacher turnover by discouraging teachers from working in schools under public scrutiny due to their low academic achievement. Recent evidence has shown that teacher turnover can have a significant detrimental effect on learning, especially in low-performing schools and those serving a high proportion of minorities and disadvantaged students (Ronfeldt, Loeb, & Wyckoff, 2013). Evidence suggests that the main mechanism behind this outcome is the organizational disruption that turnover generates. Even if teachers that leave are as effective as those who replace them, student learning could still be affected due to interruptions in social relations between teachers, which hurts staff and community cohesion, factors that are related to student engagement and achievement (Johnson, Berg, & Donaldson, 2005).

In contrast, advocates of school accountability based on standardized measures of learning maintain that accountability can result in an increase in teacher retention if they prefer teaching in lower-ranked schools. Given that this type of policy is designed to improve school performance, teachers might perceive it as a challenge that could make teaching in low-performing contexts more desirable. Similarly, principals could also manage to design and implement school-based policies that enhance teacher retention by introducing incentives and changes in the culture that make low-performing schools an attractive place to work (Dizon-Ross, 2017). Moreover, even if turnover increases this could be beneficial for school quality if lower-ability teachers are replaced by higher-ability teachers (Gjefsen & Gunnes, 2016).

Most of the empirical evidence on the effect of school accountability on teacher turnover and teacher quality composition is limited to high income countries. In the United States, studies have focused their attention on the accountability systems of Florida, North Carolina, and New York. Feng, Figlio, & Sass (2018) provide evidence that higher value-added teachers in Florida are more likely to leave schools that have received a failing grade by their accountability system. Clotfelter, Ladd, Vigdor, & Diaz (2004) used administrative data from North Carolina and explore the adverse effects on teacher retention at low-performing schools, and find that the introduction of the accountability system might have increased the share of less-experienced teachers and those with undegraduate degrees from less selective colleges. Finally, Dizon-Ross (2017) reports that the school accountability system in New York decreased teacher mobility at low performing schools and increased the overall teacher quality measured by value-added models, mainly because schools responded to accountability pressures by hiring more effective teachers.

In England, Sims (2016) uses panel data between 2010 and 2013 of primary and secondary schools to identify changes in teacher turnover in schools that were reclassified by the school inspection system from a performace category called “requires improvement” to another called “inadequate”, the lowest category. He finds that teacher turnover increased in the lowest performing category of

schools by 3.4 percentage points. His analysis does not include effects on teacher composition. In a similar study, Gjefsen & Gunnes (2016) estimate the causal effect of school accountability on teacher mobility and teacher sorting in Oslo, Norway. They report that after the program made school value added public, it generated a significant increase in the likelihood of teachers leaving the system. They also found that high-ability teachers[2] were more responsive to the reform and tended to leave the system more than other groups of teachers. However, teachers who left were also often replaced by high-ability teachers, yielding an overall positive effect.

Our work contributes to this body of literature by examining the school accountability policy in Chile. This country is characterized by a school system that provides universal per pupil vouchers to increase parental choice. While the SEP law was enacted in 2008, the accountability component of SEP was only introduced in 2012, when schools in Chile began to be ranked annually into different categories of performance. The results of the school classification are publicly available, and those schools in the lowest category not only face the potential flight of the families they serve to other relatively higher performing schools, but also the threat of closure if they do not improve within four years. We focus our attention on teachers working in low-performing schools to evaluate if they are more likely to leave their school or leave the school system than those in higher performing schools. We also estimate the potential effect for different teachers grouped according to their teaching abilities. Our results indicate that the Chilean school accountability system increased teacher turnover by increasing the likelihood of teachers leaving their school to move to another school, but not by leaving the school system. This effect is concentrated in teachers with less working experience, that teach in two or more schools, are hired through temporary contracts, and that achieve lower scores on their college admission tests. Given these results, we also tested whether accountability impacted student learning and did not find a significant effect in math achievement.

Our identification strategy relies on the fact that school rankings are defined relative to a set of variables and corresponding arbitrary thresholds for each variable. Specifically, we combine and collapse the multiple assignment variables and cutoffs used in Chile to identify low-performing schools, and then follow a traditional regression-discontinuity design (RDD) to estimate the causal effects of school accountability on teacher mobility.

In the next section, we describe the school system in Chile, the origins of the accountability program and its main features. In section 3, we present the data used for our analysis, the variables used to classify schools in the low-performing category, and how we measure teacher mobility. We also describe the methodology and test its validity. In section 4, we show the main results for all teachers and different groups of teachers. We also show other outcomes related to teacher mobility. Finally, in section 5, we present the conclusions and policy discussion.

[2] The authors constructed an ability index from teachers' grades from higher education institutions, including all universities and university colleges in Norway.

2. School accountability in Chile

The foundations of the current Chilean school system were defined in the early 1980s. One of the main features in the original design was school choice, which allowed families to choose among three types of schools: i) public schools, which are financed with government subsidies and administered by the local municipal government;[3] ii) private voucher schools, also financed with government subsidies, but administered by a private religious or secular organization; and iii) private schools, which are financed and administered privately. The argument to support this arrangement was that parents would "vote with their feet" and prefer higher quality schools, forcing poor-performing schools to either improve or go out of business (Friedman, 1955).

The government funding of public and private voucher schools was based on a per-capita scheme consisting of a flat subsidy per student enrolled (and their attendance) in the school. The per-capita subsidy only included adjustments for level (e.g. primary and secondary), modality of education (e.g. regular and special education) and school geographic location (i.e. rural and urban). Despite the expectation that the design of the system would prevent having persistently low performing schools, this did not occur. Evidence generated during the 2000's suggests that the design of the Chilean school system led to an increase in the sorting of students by abilities, creating schools with high concentrations of students with low academic achievement (Hsieh & Urquiola, 2006).

The evidence on the shortcomings of the design, combined with the pressure from secondary students demanding structural changes to the system,[4] led to the most significant structural change experienced in the Chilean school system since the 1980's. In 2008, the SEP Law was enacted. This law introduced for the first time a differentiated subsidy favoring students from disadvantaged backgrounds (low SES). The SEP law acknowledged that low SES students have greater educational needs that require extra funding and began to transfer an additional per capita subsidy (close to 50% higher than the base voucher) to schools with students classified as vulnerable[5] who attended municipal or private voucher schools that voluntarily agreed to participate in the program.[6]

[3] In 2017, the government started a process consisting in the transfer of management of public schools from municipalities to local agencies dependent of the Ministry of Education at the central level. By mid-2022, the process should be completed in all 15 regions of the country.

[4] Starting in June 2006, secondary students led multiple national protests to demand legal changes that would increase equity in education. The protests lasted around three weeks and congregated over 800,000 students in a movement popularly referred as *the penguin revolution* (O'Malley & Nelson, 2013).

[5] Vulnerable students are called "priority" students in the legal framework. This group of students include those whose parents: i) participate in the social welfare system (specifically to any of the following three programs known as *Chile Solidario, Ingreso Ético Familiar* or *Sistema Seguridades y Oportunidades*); ii) are among the poorest third according to the government's social registry of households; iii) are classified in the poorest group in the public health insurance system; or iv) if none of the above are met, the student classification is based on household income, parental education, rurality, and poverty levels in the student's municipality. After the Inclusion Law was passed in 2015, the SEP subsidy was extended to "preferential" students, reaching the lowest 60% of the national income distribution.

[6] Most schools decided to participate in the program. A year after its implementation 99% of municipal schools and 61% of private voucher schools were participating. By 2015, 100% of municipal schools and 78% of private voucher schools participated in SEP.

Along with the increase in funding, the SEP Law introduced explicit school accountability measures. For schools to receive the additional per capita subsidy for vulnerable students, they had to comply with several requirements including the signing of an agreement in which schools committed to develop and execute an improvement plan during the following four years. The agreement could be renewed if schools had spent at least 70% of the SEP additional resources, and all the expenditures had been properly reported to the relevant authorities. They also needed to comply with minimum standards of quality mainly defined by their 4th grade students' performance on standardized tests in math, science, and language over the last three years.

The quality standards were defined by a school classification that ranks schools into three categories of performance: i) *Autonomous* or high performing schools; ii) *emerging* or average performing schools, and; iii) *in-recovery*, which were low-performing schools that did not meet the minimum national standards. Although these three categories were defined in the origins of SEP, it was not until 2012 that schools started to be classified as *in-recovery*. During the period between 2008 to 2011, low-performing schools were ranked as *emerging*. It is also important to note that this classification was only valid between 2012 and 2015. Starting in 2016, the criteria to rank schools by performance changed and other dimensions of school quality were included. We focused our analysis on the *in-recovery* schools between 2012 and 2015.

During these years, the category schools received impacted the way funds were transferred to them, the degree of monitoring they were subjected to, and, eventually could lead to the revocation of the operating license granted by the Ministry of Education. While *autonomous* schools received directly the full amount of SEP funds, *emerging* schools received only one-third of their allocated SEP funding directly. The other two thirds were transferred only after the development of an improvement plan, and the execution of transfers was contingent on the correct implementation of the plan. Similarly, for *in-recovery* schools, all SEP funds were transferred in monthly installments only after the submission of the school's improvement plan. The continuation of these transfers was contingent on the correct implementation of the plan.

Accountability pressures to use resources efficiently were higher for *in-recovery* schools. According to the legislation, if these schools failed to improve their performance and move to the *emerging* category within three years, the Ministry of Education informed the school community and encouraged families to consider other schooling options, as well as facilitating transportation to another school. If the school remains in the *in-recovery* category for four years, the Ministry could revoke the school's license to operate and cut its public funding. In addition, the SEP law established that information on school performance had to be made public, which was intended to influence parental preferences. Being classified as a low-performing school could have a negative effect on future enrollments, thus affecting the amount of resources the school received through the per capita funding formula, and eventually, its ability to operate.[7]

[7] Given the change in the methodology for school classification introduced in 2016, no school classified four consecutive times as *in-recovery* between 2012 and 2015 ceased to receive public funding. Only if a school is ranked for four consecutive years ranked in the low-performing category under the newer methodology, will they no longer receive public funds.

Teachers working at *in-recovery* schools that might want to leave in response to these pressures do not face many constraints. Teachers can apply directly to private voucher schools or to any of the 346 municipalities in the country that manage public schools in their area. Paredes et al. (2013) interviewed 207 teachers in Santiago, Chile about the criteria they followed to find their first teaching position. They reported that, for the process of selecting a school, they relied on various informal social networks and it took them an average of 2.7 months to decide on a job after submitting multiple applications. In the case of principals or school managers that want to dismiss teachers, some restrictions in public schools can make it harder. The legislation only allows the dismissal of up to 5% of teachers with permanent contracts in each school per year. These types of contracts represent around 40% of all teacher contracts in public schools.

3. Data and methods

3.1 Data

To identify the effect of school accountability on teacher mobility we analyze year-to-year changes in teacher employment. Our main dependent variable is a dummy indicating if a teacher in year *t* leaves his or her school within one year. We are also interested in examining if teachers leave the school to work in another school or leave the school system. Our key independent variable is a continuous score that perfectly determines the SEP school category in a setting that allows the implementation of a regression discontinuity design. Specifically, when this score is less than zero, the school is ranked as *in-recovery* and when it is equal to or greater than zero it is either *emerging* or *autonomous*.

We combined four sets of administrative data to perform our analyses: SEP school classification databases, teacher censuses, a measurement of education quality, and school funding. The first of these contains the school performance categories for the around 9,000 schools participating in SEP each year between 2012 and 2015. It also includes the variables used for the definition of school performance.

The process of school classification follows three steps (Ley de Subvención Escolar Preferencial, 2008). First, it evaluates the three last test scores for 4th grade students on the national assessment (known as SIMCE for its acronym in Spanish),[8] it automatically assigns schools to the *emerging* category if the school has, on average, less than 20 students that took each standardized test, or if it has not participated in at least 2 out of the last 3 national assessments. Second, the remaining schools are assigned to the *in-recovery* category if they simultaneously comply with two conditions for two out of the three years considered: i) a school average is below 220 points for all the subjects assessed in 4th grade on SIMCE[9] and ii) less than 20% of its students score less than 250 in the

[8] This system is known as SIMCE for its acronym in Spanish, and every year tests all students in the 2nd, 4th, 6th, 8th, and/or 10th grades in math, language, and sciences. The SIMCE also gathers detailed information about teachers, students, and parents.

[9] The student scores for each subject tested in SIMCE follows a normal distribution with a mean of 250 and standard deviation of 50.

average of all subjects tested. Finally, another group of schools is classified as *in-recovery* if they score below the 10th percentile in the distribution of a school quality index.[10]

Table 1 shows the number of participating schools each year in the SEP classification and if they comply with the criteria to be classifiable using SIMCE data (Panel A). It also shows the total number of schools classified in each category (Panel B) and the number of schools classified following the performance indicators (Panel C). For our estimations, we are only including the schools in Panel C of Table 1. We did not include schools that were not classifiable.

Table 1: School classification by SEP each year

	Year			
	2012	2013	2014	2015
Panel A: Schools Classifiable by SIMCE	9,014	8,948	8,843	8,749
SIMCE Classifiable	3,570	3,508	3,490	3,465
Less than 20 students	3,761	3,736	3,335	2,384
Less than 2 SIMCE measures	1,683	1,704	2,018	2,900
Panel B: Classification - all schools	9,014	8,948	8,843	8,749
Autonomous	1,384	1,370	1,437	1,441
Emerging	7,430	7,389	7,347	7,237
In Recovery	200	189	59	71
Panel C: Classification - SIMCE classifiable	3,570	3,508	3,490	3,465
Autonomous	1,384	1,370	1,437	1,441
Emerging	1,986	1,949	1,994	1,953
In Recovery	200	189	59	71

Source of data is the SEP school classification databases for school years 2012 to 2015

The school-level database comprising approximately 3,500 schools each year, including performance variables, was merged with the teacher census data. This data is collected annually from principals in public and private voucher schools and includes a numerical identifier for each teacher that allowed us to trace their trajectory. We followed teachers' movements between t and t+1 between 2012 and 2016. This teacher-level database also includes information about the number of schools that a teacher works at, the type of contract, and the number of hours a week a teacher works at a given school.

To describe further differences between *in-recovery* schools and those in higher performing categories, we also complemented our main database with other databases at the school and teacher levels. Table 2 presents descriptive data for different school categories for each year between 2012 and 2015 for the variables we collected at the teacher level. Table 3 has the same format but shows data at the school level. On average, teachers at *in-recovery* schools are more likely to both move to another school and leave the school system at the end of the school year, and to work in more than one school and be hired with temporary contracts. They also are slightly more experienced,

[10] Each year a school quality index is computed based on SIMCE results (70% of the quality index) and other quality indicators such as students' approval and retention rates, parental participation at school, pedagogical innovations, adequate working conditions and teacher evaluation (30% of the index).

but do not necessarily have higher abilities. Among teachers with data on their performance on the annual teacher assessment, and achievement on college admission tests when they were applying to become teachers, (PSU for its acronym in Spanish),[11] we found that, on average, teachers who work in *in-recovery* schools have lower abilities. For example, in 2012, teachers at *in-recovery* schools are in the 48th percentile on the distribution of average math and language scores when compared with other college applicants the same year. The corresponding value for teachers in higher-ranked schools by SEP is 7 percentage points higher. The data on teacher assessment shows a similar pattern. Sixty-nine percent of teachers at *in-recovery* schools are in the two highest categories of performance, which is 9 percentage points lower than the corresponding value for *emerging* or *autonomous* schools.

The data at the school-level show that the *in-recovery* schools not only have lower results on SIMCE when compared to higher-ranked schools, but also tend to enroll fewer students and serve families with lower educational attainment that also have lower expectations regarding the future educational achievement of their children. They also receive similar funding per student from the central and local governments, but around one tenth of the school fees parents contribute.

3.2 Multivariate regression-discontinuity design

To estimate the causal effect of being classified as *in-recovery*, we follow closely the work of Elacqua, Martínez, Santos & Urbina (2016). We exploit the fact that the methodology used to classify schools in Chile is based on the schools' position relative to a set of variables and their corresponding thresholds (see section 3.1). We use a generalization of the traditional regression-discontinuity design (RDD) for the case where multiple assignment variables and cutoffs are used for treatment assignment.

[11] The PSU assessment is a college admission exam. It is comprised of four tests. Two of them are mandatory (math and language) and the other two are elective depending on the undergraduate program the applicant is pursuing.

Table 2: Teacher characteristics by SEP performance classification and year

	2012		2013		2014		2015	
	In recovery	Higher category	In recovery	Higher category	In recovery	Higher category	In recovery	Higher category
Total positions	4,270	92,919	3,965	96,403	1,270	103,047	1,608	105,220
Stays	0.73	0.78	0.72	0.80	0.78	0.82	0.76	0.81
Leaves school	0.17	0.13	0.18	0.12	0.14	0.11	0.14	0.11
Leaves system	0.11	0.09	0.09	0.07	0.08	0.07	0.10	0.07
Works in one school	0.80	0.85	0.83	0.86	0.88	0.87	0.87	0.88
Works in two or more schools	0.20	0.15	0.17	0.14	0.12	0.13	0.13	0.12
Female	0.71	0.76	0.72	0.75	0.74	0.76	0.73	0.76
Experience	14.6	13.8	13.2	12.7	12.5	12.5	13.6	12.5
Permanent contract	0.53	0.61	0.45	0.55	0.43	0.52	0.44	0.52
Contract time	32	33	33	33	34	34	34	35
Grades (percentile)	50	56	49	56	48	55	49	56
PSU (percentile)	48	55	48	54	48	54	46	55
Teacher evaluation: High	0.05	0.10	0.09	0.16	0.08	0.11	0.05	0.12
Teacher evaluation: Medium-high	0.64	0.68	0.69	0.68	0.60	0.69	0.67	0.71
Teacher evaluation: Medium-low	0.30	0.21	0.21	0.15	0.31	0.19	0.27	0.16
Teacher evaluation: Low	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01

Table 3: School characteristics by SEP performance classification and year

	Year							
	2012		2013		2014		2015	
	In recovery	Higher category	In recovery	Higher category	In recovery	Higher category	In recovery	Higher category
Total schools	196	3,366	183	3,298	55	3,420	71	3,392
Enrollment	492	675	448	662	477	656	453	670
Family annual income (US$)	6,807	10,809	7,207	11,521	7,230	11,573	8,018	12,151
Mother's schooling	9.3	11.2	9.4	11.3	9.1	11.4	9.5	11.6
Father's schooling	9.5	11.2	9.6	11.3	9.3	11.3	9.6	11.5
Expect child achieve college	0.43	0.65	0.49	0.70	0.47	0.71	0.53	0.74
Language	237	264	231	261	231	260	231	261
Math	229	258	222	253	218	251	224	255
Social Sciences	225	255	-	-	222	250	-	-
Sciences	-	-	223	251	-	-	-	-
Central government transfers per student (US$)	2,654	2,521	3,067	2,940	3,467	3,316	3,880	3,685
Local government transfers per student (US$)	162	168	181	183	185	198	211	205
Parental contribution per student (US$)	49	564	68	605	62	647	169	728

An RDD with multiple assignment variables (multivariate regression-discontinuity design or MRDD) raises challenges that are different from those identified in a traditional design, because analytical procedures for estimating treatment effects in this case are more complex and require more observations than approaches for estimating a treatment effect at a single point along a unique assignment variable. These challenges have been addressed in a series of papers (Papay et al. 2011; Reardon & Robinson, 2012; Wong et al., 2013).

Reardon and Robinson (2012) propose and discuss the merits of several estimation methods for the MRRD. We use the binding-score method. The main advantage of this approach is that it allows us to collapse scores from multiple assignment rules into a single assignment variable and therefore we can use all the observations simultaneously in the estimation. The approach also generalizes well to MRDDs with more than two assignment variables, and simplifies the analyses, avoiding the estimation of average treatment effects across multiple discontinuity frontiers. One disadvantage of this method is that pooling units from different frontiers increases the heterogeneity of the outcome at the pooled cutoff, requiring a larger bandwidth for nonparametric estimates and potentially increasing the complexity of the functional form around the cutoff (Wong et al., 2013). Despite of the latter, other studies have also chosen the binding-score strategy to obtain unbiased estimations (e.g. Robinson 2011, Reardon et al., 2010, Gill et al., 2009).

3.3 Binding-Score RD

This method is based on the construction of a new assignment variable Z (*binding-score*) that combines all the assignment variables to perfectly determine the treatment status. For example, suppose that assignment to treatment depends on two variables (R and M). Schools are assigned to a single treatment condition T if they score below both cutoffs (r_c and m_c) and to the control condition C if they score above either cutoff ($R_i \geq r_c$ or $M_i \geq m_c$). Neither of these variables individually defines treatment allocation, but we can construct a new variable Z_i, defined as the maximum between both assignment variables centered at its respective cutoff:

$$Z_i = max(R_i^c, M_i^c),$$

where $R_i^c = R_i - r_c$ y $M_i^c = M_i - m_c$, and by construction, $T_i = 1$ if $Z_i < 0$ and $T_i = 0$ if $Z_i \geq 0$

Eq. 1

Thus, the problem becomes a traditional RDD and all the standard analytical methods can be used, defining Z as the assignment variable and zero as the cutoff. Although the original assignment variables are transformed, Wong et al. (2013) show that this method estimates the same causal effect as alternative methods.

For the Chilean setting, we used the SEP ranking database to construct a binding score for each year between 2012 and 2015. First, we only considered schools that are classified by their performance indicators, excluding those that have not participated in at least two out of the last three national assessments and those with an average lower than 20 students taking each test. The

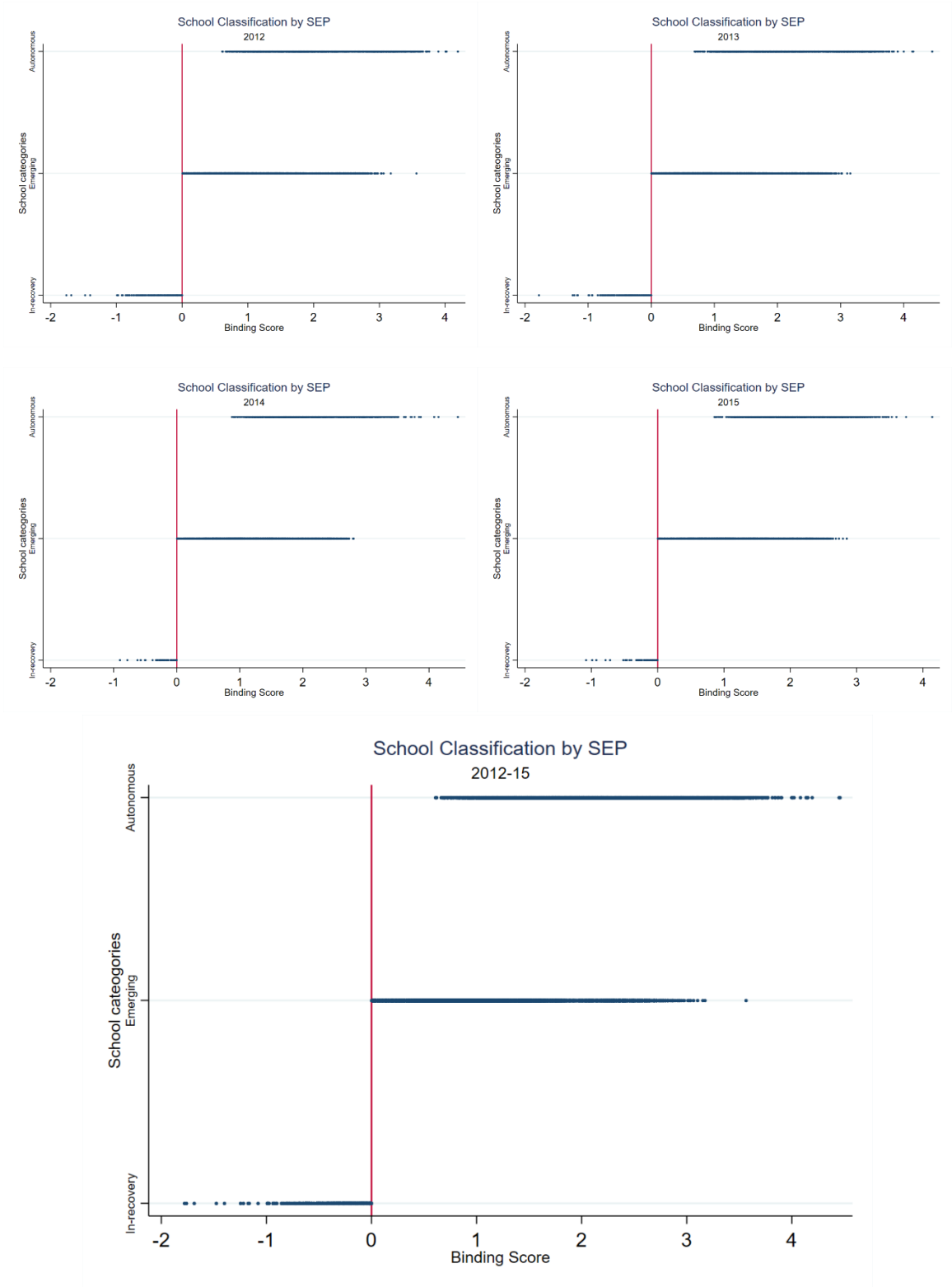
legislation defines that the schools that do not comply with these two criteria are classified as *emerging* by default and not by criteria related to their performance. Second, we combined the different assignment variables according to the criteria established in the classification methodology. Specifically, we started by standardizing all the assignment variables so that they are centered at zero. Then, we combined them by following the joint conditions defined for them.

The process of computing the binding-score for each year starts by considering that for a school to be classified as *in-recovery* it must have less than 20% of students scoring more than 250 in the average of all subjects tested by SIMCE at 4th grade, and an average lower than 220 on that same variable. These two conditions translate into the computation of three preliminary binding-scores equivalent to the maximum between these two standardized assignment variables for each of the last three years. Since this condition needs to be met for two out the last three years, the second maximum score is computed between the three-preliminary binding-scores and represents a second preliminary binding-score including all six of the conditions (two conditions each year).

To obtain the final binding-score for each year that schools were classified by SEP, the second preliminary binding-score above is combined with the last condition, which is related to a school quality index, to classify a school as *in-recovery.* Since all schools below the 10th percentile in the distribution of this index are classified as *in-recovery*, the final binding-score corresponds to the minimum between both the second preliminary binding-score and the school quality index centered at zero. The density of schools near this final binding-score each year, and for all years combined, is presented in Figure 1. Almost all schools are located within ± 1 from the cutoff separating *in-recovery* and *emerging* schools.

Using the final binding-score, we can employ a traditional RDD to estimate the effect of the program on teacher mobility. However, the credibility of this design depends critically on the inability of schools to manipulate the assignment variables, so they cannot influence their classification. In the next section we briefly discuss how the process of data-collection for the assignment variables makes it unlikely that schools can affect the SEP raking, and present two tests commonly used to provide evidence that the RDD is valid.

Figure 1: Density of schools near the binding-score cutoff for each year



Note: For each year and for all years combined, the figures plot the accountability category received by a school as a function of the underlying binding-score.

3.4 RD diagnostics

A key assumption of regression discontinuity analysis is that schools are not able to manipulate the assignment variables, thus, falling on either side of the threshold could be considered random. Some features of the SEP classification and data collection processes makes it unlikely that schools manipulated these variables. While information about the cutoffs to be classified as *in-recovery* were available in the original SEP law in 2008, these were only available for the SIMCE variables and not for the school quality index. Also, schools did not have certainty about the methodology that would be used to combine the assignment variables and might have not expected to be classified as *in-recovery* since no school received that ranking between 2008 and 2011 regardless of their performance.[12]

Additionally, the data collection process of SIMCE is closely monitored by the Ministry of Education during and after testing. The Ministry of Education hires external staff to prevent teachers and principals from having access to the tests or the classroom where the testing is being held.

We also present the results of two commonly used tests that are designed to check the validity of the RDD: i) A density test based on the work of McCrary (2008) and Cattaneo, Jansson, & Ma (2019) that attempts to identify discontinuities in the density of the assignment variables around the threshold that defines the treatment status, and ii) balance tests to show whether schools on either side of the cut-offs are observationally similar.

3.4.1. Density Test

McCrary (2008) developed a local linear density estimator to test for a discontinuity in the density of the assignment variable around the cutoff. It is computed in a two-step procedure. In the first step, a histogram of the assignment variable is constructed. In the second stage, this histogram is "smoothed" by estimating a local linear regression separately on both sides of the threshold. The test is implemented as a Wald test whose null hypothesis is that the discontinuity is zero. Under the null hypothesis of continuity, the distribution of the test is very close to a normal distribution. Cattaneo, Jansson, & Ma (2019) propose an alternative method that is fully data-driven and avoids choosing multiple tuning parameters. Based on the latter testing method, Table 4 presents the p-values associated with all assignment variables and binding-scores between 2012 and 2015. These tests suggest that there is no evidence to reject the null hypothesis of continuity in the assignment variable densities.

[12] Between 2008 and 2011 schools participating in SEP were classified as *emerging* or *autonomous*. Even though, the *in-recovery* category was defined in the law in 2008, schools could not be classified in this category until 2012.

Table 4: McCrary test implemented for all assignment variables and binding-scores (p-values)

	Year			
	2012	2013	2014	2015
Binding score	0.93	0.38	0.89	0.24
Assignment variables				
SIMCE 4th grade < 220, t-4	0.70	0.15	0.41	0.73
SIMCE 4th grade < 220, t-3	0.41	0.69	0.15	0.38
SIMCE 4th grade < 220, t-2	0.51	0.47	0.76	0.18
Proportion SIMCE 4th grade < 20%, t-4	0.85	0.14	0.54	0.49
Proportion SIMCE 4th grade < 20%, t-3	0.59	0.91	0.13	0.59
Proportion SIMCE 4th grade < 20%, t-2	0.27	0.51	0.81	0.11
School Quality Index < p10	0.46	0.99	0.34	0.32

3.4.2. Balance Tests

A second validity test frequently used to analyze the credibility of the regression discontinuity approach is to show that schools on either side of the cutoff are similar in terms of their observable characteristics (Van der Klaauw, 2008). This can be performed by testing if the relationship between the assignment variable (i.e. binding-score) and any baseline covariates is smooth in the vicinity of the discontinuity point by repeating the main analysis treating these covariates as outcomes variables.

Given that we are examining schools in various years, and that some of them are classified as *in-recovery* for more than one year, there are several ways to define the baseline. We use the values of four years before the SEP classification as our baseline. Thus, the binding-score associated with the classification in 2012 is matched with values for that same school, or the same teachers in that school, in 2008. Similarly, the binding-score associated with the classification in 2013 is matched with school or teacher variables in 2009, and so on. Following this approach, we tested for discontinuities at the cutoff between the binding-score and 11 baseline covariates, four of them at the teacher level, and the remaining seven at the school level. Table 5 shows that there are no discontinuities among the 11 variables we tested, further supporting the case that schools were not able to manipulate the assignment variables.

Table 5: Relationship between binding-score (2012-2015) with baseline covariates (2008-2011)

	Effect of classification
Pretreatment covariates	
Works in two or more schools	0.010
	(0.021)
Years of work experience	-0.418
	(0.801)
Teacher gender (=1 if female)	-0.006
	(0.025)
Weekly hours of contract	-0.47
	(0.722)
Enrollment	-0.818
	(60.743)
Family Income	15.094
	(28.457)
SIMCE math	2.038
	(2.151)
SIMCE language	-0.353
	(2.534)
Family expectations	0.001
	(0.021)
Mother schooling years	0.163
	(0.225)
Father schooling years	0.325
	(0.219)

Standard error were calculated using clusters at the school-year level and are shown in parenthesis.

To define the baseline, the binding-score associated with the classification in 2012 (2013, 2014, 2015) is matched with values for that same school, or the same teachers in that school, in 2008 (2009, 2010, 2011).

3.5 Empirical Strategy

We are interested in estimating the effect that working in an *in-recovery* school has on the probability of teachers leaving their school. We started by defining a set of dummy variables as outcomes that are equal to one when the teachers move or leave the school, either by moving to another school or by leaving the school system entirely. The binary nature of our dependent variables introduces some challenges that can be addressed in different ways. The simplest approach of using a binary variable as the dependent variable has been questioned by Xu (2017),

who argues that the commonly used RD approach proposed by Calonico, Cattaneo, & Titiunik (2014) fails to determine an optimal bandwidth and compute robust standard errors in this case.

One common practice to circumvent this issue with binary outcomes is to first aggregate them over different bins along the binding-score so the outcomes are transformed to fractions that can be treated as continuous. However, this strategy introduces an additional tuning parameter: the bin size for aggregation. The automatic method proposed by Calonico, Cattaneo, & Titiunik (2015) can be used, but is not clear how this affects the final RD estimate.

Furthermore, when we attempt to estimate the separate effect on moving to another school and leaving the school system, additional issues arise when working with categorical variables. A standard practice to address them is to first generate a binary outcome for each category (against the chosen baseline category) and then applying the standard RD methods for each binary outcome. However, besides the issue of determining the bin size, this strategy focuses on each category in isolation, ignoring the correlation among responses, and thus does not support simultaneous inference of treatment effects across categories.

Xu (2017) proposes a nonparametric strategy to perform analyses under a sharp RDD with a categorical outcome. He extends the multinomial logit model to allow the estimation of optimal bandwidths around the cutoff and robust confidence intervals for the treatment effects. However, his approach also has at least two relevant limitations. First, it does not allow the computation of clustered standard errors, which might affect the inference of our estimations at the teacher-level. Second, if covariates are included, there are no adjustments to prevent potential inconsistency or invalid inference in RD estimators.

Since no approach is necessarily superior, we employ various strategies to analyze teacher mobility. We begin by using a non-parametric approach and pooling the data from all years between 2012 and 2015. This model fits a kernel-weighted linear regression for teacher observations in an interval around the cutoff defined by the bandwidth h. Specifically, the base model is as follows:

$$Y_{t+1} = \alpha + \tau \cdot IR_t + \beta_1 \cdot (Z_t - c) + \beta_1 \cdot IR_t \cdot (Z_t - c) + \epsilon$$

$$\text{where } c - h \leq Z_t \leq c + h$$

(Eq. 2)

Teacher mobility is represented by Y_{t+1}, which examines if the teacher stayed or left the school in the year immediately following the school classification. IR_t is a dummy variable that takes the value of one when the schools are classified as *in-recovery* in year t and is zero otherwise. Variable Z_t corresponds to the binding-score, and c is the cutoff. Is important to note that the treatment condition is defined by $IR_t = 0$ if $Z_t \geq c$ and $IR_t = 1$ if $Z_t < c$. Our binding-scores are computed so that the cutoff that defines treatment status is zero ($c = 0$ for all binding-scores defines the cutoff that divides *in-recovery* and *emerging* schools). The impact that working at a school classified as *in-recovery* has on teacher mobility is captured by the coefficient τ.

A key decision under the nonparametric approach above is the bandwidth size. The bandwidth defines the weight assigned to each observation. As the bandwidth gets smaller, the observations

close to the cutoff receive more weight in the estimation. We present as our main specification the results associated with the optimal bandwidth proposed by Calonico, Cattaneo, Farrell, & Titiunik (2017), based on the approach known as Mean Squared Error (MSE) optimal bandwidth for the RD treatment effect estimator. To test the robustness of our results, we estimate models with various bandwidths, using three factors that expand the optimal bandwidth to both sides (1.75; 1.50; and 1.25), and three factors that contracts it at both sides (0.75; 0.50; 0.25).

We follow this strategy to estimate the impact that working at an *in-recovery* school has on the five dummy variables that measure teacher mobility. The first three use as the base category (equal to zero) when the teacher stays in the same school in $t + 1$ and 1 when the teacher: i) moves to another school or leaves the school system; ii) moves to another school; iii) leaves the school system. The fourth variable takes the value of one when the teacher moves to another school, as our second dummy variable, but is equal to zero when the teacher stays in the school *or* leaves the school system. The fifth variable is equal to one when the teacher leaves the school system, as our third binary variable, but in this case the base category (that takes the value of zero), corresponds to the case when the teacher stays in the same school or moves to another school. The two latter variables are included in the analysis to check if our results are robust to the choice of the base category.

For each of our five binary variables we estimate five different specifications. For the first and second, we use the binary variable as the outcome as defined above and perform the estimations as defined by equation 2. The only difference is that for the second specification we include a set of covariates including year fixed-effects, family income, and an indicator showing if the teachers works simultaneously in two or more schools. In the specifications 3 to 5 we use the same specification as in the first one, but we use a linearized version of the dummy variable using different bin sizes to compute the average in the probability of departure from their school. The bin size is determined based on the automatic method proposed by Calonico, Cattaneo, & Titiunik (2015). We also present computed averages using bins that are two and one-half times the optimal bin size suggested by these authors.

We implement the estimations above following the work of Calonico, Cattaneo, Farrell, & Titiunik (2017), who propose a method that allows the inclusion of continuous, discrete, and mixed additional regressors. This is performed through a procedure that adjusts the model to avoid inconsistent RD estimators, that continues to compute the MSE-optimal bandwidth, and that allows for a valid asymptotic inference including heteroskedastic and clustered data.

Additionally, to complement the analysis based on the five binary outcomes above, in the Appendix 1 we also present the results following the strategy proposed by Xu (2017) to estimate the effect of school accountability on teacher mobility in $t + 1$ measured by a categorical variable comprised of three categories: i) leaves the school system; ii) leaves the school but not the school system; iii) stays in the same school. We use the latter as the base category.

Finally, for all of our outcome variables, we perform a *placebo* test to check whether there are any baseline differences in teacher turnover between schools on either side of the threshold. We estimate the exact same regression models we used to identify treatment effects on our five binary outcomes but using the teacher mobility of previous years. To perform this exercise, we follow the

same data matching strategy we used to show the absence of unbalances at baseline in the 11 covariates presented in Table 5. Specifically, the binding-score associated with the classification in 2012 is matched with teacher mobility for that same school in 2008. Likewise, the binding-score associated with the classification in 2013 is matched with the teacher mobility in the same school in 2009, and so on.

4. Results

This section is organized to display four sets of results. First, we present our main outcome including robustness checks and some exercises that helps to understand the policy effect on teacher mobility. Second, we show results associated with estimations testing whether families also reacted to accountability pressures by leaving *in-recovery* schools. Third, we present the results of an exercise examining the effects of school accountability on new hires. Finally, we provide empirical evidence on the potential influence of school accountability threats on student achievement.

4.1 Results for teacher mobility

4.1.1 Main outcome

Table 6 presents the effect that working at an *in-recovery* school has on teacher mobility measured by the five binary variables described in section 3.4. For ease of interpretation, the coefficients were multiplied by negative one, so a positive value corresponds to an increase in the likelihood of leaving the school. Column (1) shows the effects following Equation 1 for a dummy variable equal to one when the teacher moves to another school or leaves the school system the year after the school was classified as *in-recovery*, and zero when he or she stays in the same school. The results show that teachers working at *in-recovery* schools are more likely to leave their schools by around 4 percentage points. This result is consistent across the five specifications (see section 3.4).

Column (2) reports the effect on the likelihood that the teacher will leave the *in-recovery* school by moving to another school, and Column (3) shows the effect on the likelihood that the teacher will leave the school system. The computation of both effects is based on binary variables with the same base category: the teacher stays in the same school the following year. The results indicate that the program affected the likelihood of teachers moving to another school, but not leaving the system. A graphical representation of the results from column (1) to (3) in Table 6 are presented in Figure 2.

Our results are consistent when we change the base category. Column (4) presents the results for the outcome of teachers moving to another school but using as the base category teachers that stay in the same school or leave the school system. Similarly, column (5) shows the effect on teachers moving out of the school system. The magnitude of the effect on moving to another school is

between 3.7 to 4.1 percentage points which represents around 25% of the baseline. We do not find any effect on the likelihood of leaving the school system.

Table 6: RD estimates of the effect of working at an *in-recovery* school on teacher mobility

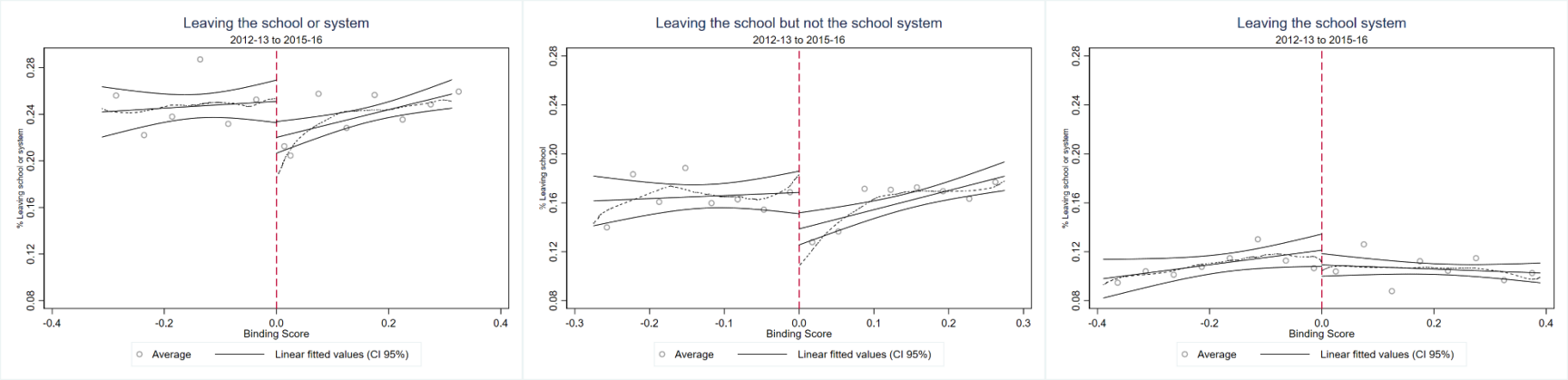
	Binary Outcomes				
	Leaves school or system = 1; Stays = 0	Leaves school = 1; Stays = 0	Leaves system = 1; Stays = 0	Leaves school = 1; Stays or leaves the system = 0	Leaves system = 1; Stays or leaves the school = 0
	(1)	(2)	(3)	(4)	(5)
Dichotomized variable					
Treatment effect	0.041*	0.041**	0.013	0.037**	0.008
	(0.022)	(0.019)	(0.014)	(0.017)	(0.012)
Bandwidth	0.31	0.27	0.39	0.27	0.4
Number of Obs: Left - Right	7021-17176	5709-13446	6994-19175	6255-14418	8310-23704
Dichotomized variable including covariates§					
Treatment effect	0.033*	0.042**	0.007	0.038**	0.002
	(0.02)	(0.018)	(0.015)	(0.016)	(0.012)
Bandwidth	0.31	0.24	0.35	0.23	0.35
Number of Obs: Left - Right	6793-16476	5096-11562	6354-16087	5479-12100	7482-18848
Linearized variable, bins width = 0.0005					
Treatment effect	0.041**	0.042***	0.013	0.037***	0.008
	(0.016)	(0.015)	(0.012)	(0.013)	(0.01)
Bandwidth	0.31	0.27	0.39	0.27	0.4
Number of Obs: Left - Right	7021-17176	6317-14806	8214-22510	6255-14418	8310-23704
Linearized variable, bins width = 0.0010					
Treatment effect	0.041***	0.042***	0.014	0.037***	0.008
	(0.014)	(0.013)	(0.01)	(0.012)	(0.008)
Bandwidth	0.31	0.27	0.39	0.27	0.4
Number of Obs: Left - Right	7021-17176	6317-14806	8214-22510	6255-14418	8310-23704
Linearized variable, bins width = 0.0015					
Treatment effect	0.041***	0.04***	0.012	0.037***	0.008
	(0.009)	(0.009)	(0.008)	(0.009)	(0.007)
Bandwidth	0.31	0.27	0.39	0.27	0.4
Number of Obs: Left - Right	7021-17176	6317-14806	8214-22510	6255-14418	8310-23704

Standard error were calculated using clusters at the school-year level and are shown in parenthesis

§: Covariates include years fixed effects, a dummy indicating whether the teacher works in two or more schools, teacher experience, gender, type of contract (=1 if permanent) and contract hours. We also included variables at school level: enrollment, average of family income, and SIMCE math.

∗ significant at 10%; ∗∗ significant at 5%; ∗∗∗ significant at 1%.

Figure 2: Graphical RD estimates of the effect of working at an *in-recovery* school on teacher mobility



Note: Each dot represents the average within 0.05 units of the binding-score. The x-axes show the schools' average binding-score and a horizontal red line is at the cutoff (equal zero). The y-axes show the average teacher mobility for each of the three teacher mobility variables analyzed: (i) moves to another school or leaves the school system; (ii) moves to another school, and; (iii) leaves the system. The straight lines correspond to the local estimation based on a polynomial equal 1 within the optimal bandwidth and the curve lines represent the 95% confidence intervals.

4.1.2 Robustness checks for main outcome

We perform two robustness checks. First, since the nonparametric approach can depend on the bandwidth size, we replicate the estimations reported in column (1) in Table 6 using six other bandwidths. Three of these are bigger than the preferred bandwidth (1.75; 1.50; and 1.25 times) and three of them smaller (0.75; 0.50; and 0.25 times). The results of this exercise are presented in Table 7 and they show that the effect on teachers moving out of *in-recovery* schools seems to be robust to the bandwidth size.

Table 7: Robustness of the teacher mobility estimates to the size of the bandwidth

	Outcomes				
	Leaves school or system = 1; Stays = 0	Leaves school = 1; Stays = 0	Leaves system = 1; Stays = 0	Leaves school = 1; Stays or leaves the system = 0	Leaves system = 1; Stays or leaves the school = 0
	(1)	(2)	(3)	(4)	(5)
Bandwidths					
0.25 x MSE Op	0.069*	0.067*	0.022	0.061*	0.013
	(0.041)	(0.038)	(0.025)	(0.035)	(0.021)
0.50 x MSE Op	0.055*	0.069***	0.002	0.063***	-0.004
	(0.029)	(0.026)	(0.019)	(0.024)	(0.016)
0.75 x MSE Op	0.040	0.051**	0.009	0.047**	0.004
	(0.024)	(0.022)	(0.016)	(0.02)	(0.013)
MSE Op	0.041*	0.041**	0.013	0.037**	0.008
	(0.022)	(0.019)	(0.014)	(0.017)	(0.012)
1.25 x MSE Op	0.039**	0.037**	0.015	0.033**	0.01
	(0.019)	(0.017)	(0.013)	(0.016)	(0.011)
1.5 x MSE Op	0.034*	0.035**	0.013	0.03**	0.008
	(0.018)	(0.016)	(0.012)	(0.014)	(0.01)
1.75 x MSE Op	0.025	0.028*	0.01	0.024*	0.007
	(0.017)	(0.015)	(0.012)	(0.014)	(0.01)

Standard error were calculated using clusters at the school-year level and are shown in parenthesis

∗ significant at 10%; ∗∗ significant at 5%; ∗∗∗ significant at 1%.

The second robustness check we perform is a placebo test using the same five binary variables measuring teacher mobility but using values for four years before schools received the *in-recovery* status. The results are displayed in Table 8, and they show that none of the placebo coefficients are statistically significant.

Table 8: RD estimates of the effect of working at an *in-recovery* school on pre-treatment teacher mobility

	Outcomes				
	Leaves school or system = 1; Stays = 0	Leaves school = 1; Stays = 0	Leaves system = 1; Stays = 0	Leaves school = 1; Stays or leaves the system = 0	Leaves system = 1; Stays or leaves the school = 0
	(1)	(2)	(3)	(4)	(5)
Dichotomized variable					
Treatment effect	0.002	0.002	0.003	0.002	0.003
	(0.024)	(0.02)	(0.014)	(0.017)	(0.011)
Bandwidth	0.29	0.26	0.34	0.26	0.37
Number of Obs: Left - Right	6674-15104	5648-12074	6692-15784	6192-13257	8085-19968
Dichotomized variable including covariates[§]					
Treatment effect	-0.001	0.006	0.000	0.004	0.000
	(0.023)	(0.018)	(0.014)	(0.016)	(0.011)
Bandwidth	0.29	0.25	0.35	0.25	0.38
Number of Obs: Left - Right	6539-14393	5274-11215	6684-15925	5789-12361	8138-20471
Linearized variable, bins width = 0.0005					
Treatment effect	0.002	0.002	0.003	0.002	0.003
	(0.021)	(0.018)	(0.012)	(0.016)	(0.009)
Bandwidth	0.29	0.26	0.34	0.26	0.37
Number of Obs: Left - Right	6674-15104	6192-13283	7640-17758	6192-13257	8085-19968
Linearized variable, bins width = 0.0010					
Treatment effect	0.002	0.003	0.003	0.002	0.003
	(0.018)	(0.015)	(0.011)	(0.012)	(0.008)
Bandwidth	0.29	0.26	0.34	0.26	0.37
Number of Obs: Left - Right	6674-15104	6192-13283	7640-17758	6192-13257	8085-19968
Linearized variable, bins width = 0.0015					
Treatment effect	0.002	0.001	0.003	0.001	0.003
	(0.015)	(0.013)	(0.008)	(0.011)	(0.006)
Bandwidth	0.29	0.26	0.34	0.26	0.37
Number of Obs: Left - Right	6674-15104	6192-13283	7640-17758	6192-13257	8085-19968

Standard error were calculated using clusters at the school-year level and are shown in parenthesis

§: Covariates include years fixed effects, a dummy indicating whether the teacher works in two or more schools, teacher experience, gender, type of contract (=1 if permanent) and contract hours. We also included variables at school level: enrollment, average of family income, and SIMCE math.

4.1.3 Results by year of SEP policy

We also estimate the effect of school accountability on teacher mobility separately and independently for each year the policy was implemented. Just as we did for the pooled data, we use five different dependent variables that measure teacher mobility. Table 9 display the results showing that the increased likelihood of leaving the *in-recovery* school is concentrated in the first year of the program. For each year of the program, we also failed to identify changes in the likelihood of leaving the school system due to the accountability component of SEP.

Table 9: RD estimates of the effect of leaving an *in-recovery* school by year of program

	Outcomes				
	Leaves school or system = 1; Stays = 0	Leaves school = 1; Stays = 0	Leaves system = 1; Stays = 0	Leaves school = 1; Stays or leaves the system = 0	Leaves system = 1; Stays or leaves the school = 0
	(1)	(2)	(3)	(4)	(5)
Year of program					
First year - 2012	0.042	0.074**	0.013	0.070**	0.004
	(0.033)	(0.038)	(0.029)	(0.034)	(0.023)
Second year - 2013	0.028	0.054	-0.015	0.054*	-0.019
	(0.039)	(0.033)	(0.028)	(0.03)	(0.023)
Third year - 2014	0.043	0.001	0.048	-0.004	0.043
	(0.066)	(0.049)	(0.042)	(0.043)	(0.037)
Fourth year - 2015	0.008	0.037	-0.024	0.032	-0.025
	(0.049)	(0.045)	(0.033)	(0.04)	(0.028)

Standard error were calculated using clusters at the school-year level and are shown in parenthesis

Number of observations to the left of threshold within the optimal bandwith range from 831 for 2014 to 3,001 in 2012. Number of observations to the right of the threshold within the optimal bandwith range from 2,458 in 2014 to 6,834 in 2012

$*$ significant at 10%; $**$ significant at 5%; $***$ significant at 1%.

4.1.4 Results for schools that were classified more than once as *In-Recovery*

In this subsection we provide evidence suggesting that schools classified multiple times as *in-recovery* do not face a higher departure of teachers compared to one time *in-recovery* schools. Analyzing this issue is challenging given the low number of schools classified three or more times as *in-recovery*. Thirty-one schools were classified three times as *in-recovery* between 2012 and 2015 and only 19 were given the same classification four times. Estimations based on a low quantity of schools might not provide enough statistical power to identify an effect of the ranking on mobility.

For the reason above, we focused this analysis on the schools labeled as *in-recovery* during 2013. Out of the 183 schools classified as *in-recovery* in 2013, there are 107 that were also classified as *in-recovery* in 2012. Using the subsample of schools classified as *in-recovery* in 2013 and the corresponding binding score, we are able to identify the effect of being classified as in-recovery twice (treatment) in comparison to those schools that were classified as in-recovery only once during 2013 (control).

Table 10 displays the results on teacher mobility of working at a school classified as in-recovery twice compared to working in a school that has been in the same category for only one year. None of the five mobility variables show a significant effect. While we are not able to analyze teachers in schools classified as *in-recovery* for three or four times, the combination of results from the former subsection showing no effects on teacher mobility for years 2014 and 2015 with those in this subsection showing no effect on those teachers working in schools classified as *in-recovery* twice, still suggest that the number of times schools was classified as *in-recovery* is unrelated to teacher mobility. The effect of accountability pressure on teacher mobility seems to occur only after the first year a school is ranked as low performing.

Table 10: RD estimates of the effect of leaving an *in-recovery* school year of program

	Outcomes				
	Leaves school or system = 1; Stays = 0	Leaves school = 1; Stays = 0	Leaves system = 1; Stays = 0	Leaves school = 1; Stays or leaves the system = 0	Leaves system = 1; Stays or leaves the school = 0
	(1)	(2)	(3)	(4)	(5)
Dichotomized variable					
Treatment effect	-0.007	0.006	-0.017	0.007	-0.02
	(0.091)	(0.086)	(0.047)	(0.077)	(0.039)
Bandwidth	0.2	0.18	0.17	0.18	0.16
Number of Obs: Left - Right	729-867	642-747	564-638	695-832	600-705
Dichotomized variable including covariates§					
Treatment effect	-0.002	0.026	-0.031	0.026	-0.033
	(0.085)	(0.084)	(0.038)	(0.076)	(0.031)
Bandwidth	0.19	0.18	0.17	0.18	0.16
Number of Obs: Left - Right	699-811	615-690	548-603	679-771	584-684

Standard error were calculated using clusters at the school-year level and are shown in parenthesis

§: Covariates include years fixed effects, a dummy indicating whether the teacher works in two or more schools, teacher experience, gender, type of contract (=1 if permanent) and contract hours. We also included variables at school level: enrollment, average of family income, and SIMCE math.

$*$ significant at 10%; $**$ significant at 5%; $***$ significant at 1%.

4.1.5 Heterogeneous effects

We now present the effect of the school accountability policy on different types of teachers. We used five variables to group teachers. The first two variables are related to the working conditions and the last three variables are proxies of teacher ability. The first corresponds to the type of contract teachers have, which could be either permanent, when teachers have tenure, or temporary, if teachers are hired for up to two years. The second separates teachers into two groups according to the number of schools they are working in *t*; one school and two or more schools. The third variable is the teachers' average scores on standardized tests of math and language that are used for college admission. We were able to merge these scores for around 22% of all teachers. These teachers were enrolled in a teaching undergraduate program after 2003. Using this data, we define two groups, those above the 50^{th} percentile in the distribution of scores on the admissions test the year they applied to college, and those below that threshold. For those teachers with scores for more than one year, we used the most recent score.

Fourth, we use the teacher's ranking on the national teacher assessment performed annually on a sample of public-school teachers. This assessment classifies teachers into four performance categories based on their scores on the quality of their classes, learning materials, and evaluations conducted by peers and self-assessments. We managed to identify all teachers that were assessed between 2008 and 2015 and used their most up to date classification to separate them into two groups. Those in the highest two categories of performance (around 75% of assessed teachers) and

those in the lowest two (remaining 25%). Approximately 35% of teachers in our study also had data from the national teacher assessment.

Finally, we use teacher experience as a third variable to group teachers. This variable is reported for all teachers. We use three years of working experience as a threshold to define two groups of teachers. Unexperienced or novice teachers that have less than three years of working experience, and more experienced teachers comprised of those with three or more years of experience (Kane et al., 2008).

We define a total of ten groups of teachers following the criteria described above. For each of these groups we estimated separately the effect of working at an *in-recovery* school on teacher mobility using the same five binary variables and specifications presented in Tables 6 to 10. Table 11 presents the results of this analysis and shows that teachers leaving *in-recovery* schools are more likely to teach in two or more schools, be hired through temporary contracts, have less working experience, and have lower scores on their college admission tests.

Regarding the variables proxying teacher abilities, our estimations indicate that the likelihood of moving to another school the year after the school is ranked in the lowest performance category increases by around 10 percentage points for teachers with less than 3 years of working experience, and by around 13 percentage points for teachers scoring lower on the college admission exam. When we tested both groups of teachers separated by performance on the national teacher assessment for teachers in public schools, we do not find any significant effects.

Given that we only have data for performance on the college admission tests for teachers that started teaching programs after 2003, the results taken together suggest that younger teachers that are less qualified are more likely to leave schools ranked in the *in-recovery* category. However, these results should be considered as suggestive since the sample of teachers with college admission data is more limited and the predictive power of this score on teacher effectiveness is weaker than the one associated with the teacher assessment scores (San Martín, Rivero, Bascopé, & Hurtado, 2013).

Table 11: RD estimates of the effect of working at an *in-recovery* school on teacher mobility for different groups of teachers

	Outcomes				
	Leaves school or system = 1; Stays = 0	Leaves school = 1; Stays = 0	Leaves system = 1; Stays = 0	Leaves school = 1; Stays or leaves the system = 0	Leaves system = 1; Stays or leaves the school = 0
	(1)	(2)	(3)	(4)	(5)
Teacher Groups					
Permanent contract	0.007	0.021	-0.001	0.021	0.000
	(0.024)	(0.017)	(0.018)	(0.019)	(0.016)
Short term contract	0.063**	0.055**	0.028	0.066**	0.015
	(0.026)	(0.024)	(0.023)	(0.027)	(0.017)
Works in one school	0.023	0.027	0.009	0.028	0.006
	(0.023)	(0.018)	(0.016)	(0.02)	(0.014)
Works in two or more schools	0.128***	0.101**	0.041	0.112**	0.017
	(0.049)	(0.048)	(0.026)	(0.05)	(0.017)
Unexperienced teacher (<= 3 years)	0.106**	0.093**	0.034	0.11***	-0.042
	(0.043)	(0.037)	(0.038)	(0.042)	(0.077)
Experienced teacher (> 3 years)	0.025	0.023	0.012	0.025	0.034
	(0.022)	(0.017)	(0.014)	(0.019)	(0.071)
Lower teacher evaluation category	0.028	0.028	0.003	0.033	0.023
	(0.044)	(0.038)	(0.033)	(0.046)	(0.113)
Higher teacher evaluation category	0.021	0.035	-0.006	0.035	0.021
	(0.026)	(0.023)	(0.016)	(0.024)	(0.096)
Lower college entry score	0.126**	0.129***	0.022	0.140***	0.012
	(0.054)	(0.044)	(0.044)	(0.049)	(0.105)
Higher college entry score	0.073	0.027	0.075	0.044	-0.028
	(0.06)	(0.052)	(0.05)	(0.058)	(0.093)

Standard error were calculated using clusters at the school-year level and are shown in parenthesis

Number of observations to the left of threshold within the optimal bandwith range from 598 for the group "Higher college entry score" to 5,549 for the group "Experienced teacher". Number of observations to the right of the threshold within the optimal bandwith range from 1,778 for the group "Higher college entry score" to 13,360 for the group "Experienced teacher"

∗ significant at 10%; ∗∗ significant at 5%; ∗∗∗ significant at 1%.

One potential reason that might explain the increased likelihood of teachers leaving *in-recovery* schools is that families are also reacting to the low school ranking by moving to another school or leaving the school system. To test for this possibility, we performed two exercises. First, we identify students at the primary level that moved out from their schools to attend another school or leave the school system the following year. We created five mobility variables for students in grades 1 to 7 between the school years 2012 and 2015. These variables are equivalent to the five teacher mobility variables we analyzed in section 4.1. To estimate the causal effect of attending an *in-recovery* school on student mobility we followed the same methodology described in section 3.

Second, using school enrollment data for each year between 2012 and 2016, we computed the difference between enrollment the year following the receipt of the SEP classification and the year the school received it. We computed this difference for enrollment in preschool, primary, secondary, and all levels. Then we pooled the data to estimate the effect of school accountability on enrollment following the same approach we used for previous sections.

Table 12 present the results for both analyses. Panel A shows the results for the outcomes of student mobility and Panel B for school enrollment for different educational levels. Both show the same result that school accountability in Chile had no effect on family decisions to move out of *in-recovery* schools, which suggests that teacher departure from these schools seems not to be in response to a decreased student enrollment.

Table 12: RD estimates of the effect of being classified as *in-recovery* on student mobility (Panel A) and school enrollment (Panel B)

PANEL A: Students mobility	Outcomes				
	Leaves school or system = 1; Stays = 0	Leaves school = 1; Stays = 0	Leaves system = 1; Stays = 0	Leaves school = 1; Stays or leaves the system = 0	Leaves system = 1; Stays or leaves the school = 0
	(1)	(2)	(3)	(4)	(5)
Dichotomized variable					
Treatment effect	-0.01	-0.008	-0.003	-0.008	-0.003
	(0.012)	(0.011)	(0.005)	(0.01)	(0.004)
Bandwidth	0.27	0.27	0.29	0.27	0.3
Number of Obs: Left - Right	73,340-176,181	71,221-172,264	63,028-154,896	74,321-177,928	80,138-196,032
Dichotomized variable including covariates§					
Treatment effect	-0.012	-0.01	-0.004	-0.009	-0.003
	(0.012)	(0.011)	(0.005)	(0.01)	(0.004)
Bandwidth	0.27	0.27	0.26	0.27	0.27
Number of Obs: Left - Right	70,898-169,177	69,661-165,226	57,461-137,716	72,263-174,313	72,263-173,184
PANEL B: School Enrollment	Enrollment at				
	Preschool	Primary	Secondary	Total	
	(1)	(2)	(3)	(4)	
Dichotomized variable					
Treatment effect	-1.718	1.026	0.112	0.46	
	(1.852)	(4.456)	(1.688)	(5.573)	
Bandwidth	0.38	0.38	0.45	0.4	
Number of Obs: Left - Right	356-935	354-920	386-1,164	364-1,003	
Dichotomized variable including covariates§					
Treatment effect	-2.01	0.497	2.263	-0.923	
	(1.927)	(3.866)	(1.844)	(5.109)	
Bandwidth	0.36	0.50	0.29	0.48	
Number of Obs: Left - Right	336-831	397-1,341	283-641	393-1,267	

Standard error were calculated using clusters at the school-year level and are shown in parenthesis

§: Covariates for estimations in Panel A include years fixed effects, student age, and student gender. We also included school-level variables: enrollment, average of family income, and SIMCE math. Covariates for estimations in Panel B include years fixed effects, and school-level variables: average of family income, and SIMCE math.

$*$ significant at 10%; $**$ significant at 5%; $***$ significant at 1%.

4.3 Are *in-recovery* schools responding by hiring new teachers?

To study if *in-recovery* schools reacted attempting to replace teachers leaving by hiring new teachers, we estimated the effect of the accountability program on two dummy variables that identify new teachers at each school the year immediately following the school classification. The first of these identifies new teachers regardless of their work experience, (i.e. they might have

worked at another school in the past), and the second identifies new teachers that are working for the first time as teachers.

The results are reported in Table 13. Just as we did in our analysis identifying teacher departures, we estimated the effect of school accountability on different specifications and definitions for the binary outcomes. The results for all specifications show no effect on the hiring of new teachers for *in-recovery* schools. In other words, the school accountability system in Chile seems not to have incentivized the hiring of new teachers at *in-recovery* schools, even though a group of teachers are exiting these schools.

Table 13: RD estimates of the effect of being classified as *in-recovery* on teacher hiring

	Outcomes	
	New with the school = 1; Not new with the school = 0	New with the system = 1; Not new with the system = 0
	(1)	(2)
Dichotomized variable		
Treatment effect	-0.002	-0.005
	(0.02)	(0.011)
Bandwidth	0.37	0.36
Number of Obs: Left - Right	8,050-21,282	7,926-20,658
Dichotomized variable including covariates[§]		
Treatment effect	0.006	-0.002
	(0.015)	(0.01)
Bandwidth	0.46	0.35
Number of Obs: Left - Right	8,919-28,943	7,556-19,201

Standard error were calculated using clusters at the school-year level and are shown in parenthesis

§: Covariates include years fixed effects, a dummy indicating whether the teacher works in two or more schools, teacher experience, gender, type of contract (=1 if permanent) and contract hours. We also included variables at school level: enrollment, average of family income, and SIMCE math

4.4 School accountability and student achievement

As a final exercise, we test whether the Chilean school accountability system impacted student learning. It appears that less-effective teachers are leaving *in-recovery* schools and, even though they are not being replaced, it still may have a positive or negative impact on student learning. To test this, we focused on the 2012 cohort of schools for which our results from previous sections suggest the teachers reacted the most to the accountability system.

We estimate the effect on SIMCE math in grades 4, 6, and 8 using school-level data and by following our regression discontinuity approach based on the 2012 binding score described in

section 3.3. We considered as treated schools those that were classified as *in-recovery* in 2012, and as control schools those in other categories of the SEP classification that same year. [13]

We estimate the effect in seven different math outcomes for all grades evaluated by SIMCE. The first six outcomes correspond to each SIMCE math score between 2011 to 2016. We included the results for math 2011 as a pretreatment variable. The math 2012 represents the short-term effect of the program, and the following years represent longer term effects. The seventh outcome corresponds to a measure of school value added. The value-added model was used following the 2011 cohort of 4th grade students that were reevaluated in 2013 when they were in 6th grade, and in 2015 when they reached 8th grade. The value-added model was estimated using student-level data, following a school fixed effects model, and including as control variables age and gender of students, mother's schooling, and family's income at baseline.

Table 14 shows the effect of accountability on math achievement for all grades assessed by SIMCE between 2011 and 2016, including outcomes of value added between grades 4 and 6, and 4 and 8, as described above. The results show that the coefficients are not statistically significant suggesting that the Chilean school accountability system was not an effective policy to improve student learning outcomes measured by standardized tests. Conversely, accountability also did not appear to have a negative impact on learning.

Table 14: RD estimates of the effect of being classified as *in-recovery* on student achievement on math

	SIMCE Math						
	2011	2012	2013	2014	2015	2016	Value added Baseline 4th grade 2011
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
4th grade							
Treatment effect	0.033	0.19	-0.008	0.057	-0.074	-0.216	-
	(0.180)	(0.223)	(0.219)	(0.243)	(0.212)	(0.233)	-
Bandwidth	0.400	0.380	0.390	0.340	0.340	0.350	-
Number of Obs: Left - Right	134-321	129-293	132-304	112-253	113-253	112-264	-
6th grade							
Treatment effect	-	-	0.055	-0.165	0.01	-0.003	0.16
	-	-	(0.116)	(0.147)	(0.17)	(0.165)	(0.231)
Bandwidth	-	-	0.500	0.330	0.380	0.340	0.350
Number of Obs: Left - Right	-	-	151-419	111-243	127-291	109-252	99-224
8th grade							
Treatment effect	-0.033	-	0.13	0.024	-0.036	-	0.021
	(0.144)	-	(0.135)	(0.107)	(0.165)	-	(0.274)
Bandwidth	0.35	-	0.38	0.51	0.37	-	0.3
Number of Obs: Left - Right	114-253	-	131-294	143-428	122-279	-	72-159

Standard error were calculated using clusters at the school-year level and are shown in parenthesis

[13] In Appendix 2, we estimate the same models from section 4.4 but excluding schools that were not classified as in-recovery in 2012 and classified as in-recovery at least once between 2013 and 2015. The results do not differ to those presented in this section.

5. Discussion

There are conflicting views in the literature on the impact of accountability on teacher mobility. Critics are concerned that the negative stigma of working at a low-performing school will drive out the best teachers of the most disadvantaged schools. Advocates counter that accountability will either create incentives for schools to retain their highest performing teachers to avoid negative sanctions or dismiss the least effective teachers. This argument is predicated on the assumption that low performing schools will be able to replace the lowest performing teachers with better teachers.

By exploiting a set of arbitrary rules used to classify low-performing schools, our work is the first to evaluate the effect of school accountability on teacher mobility in a developing country. In line with what most of the empirical literature on the subject reports, we found that classifying a school as low-performing increases teacher departure the following year, and for our case, this is not offset by the hiring of new teachers.

Using different empirical strategies, we show that after the policy was introduced, teachers were more likely to leave *in-recovery* schools by 3.3 to 4.1 percentage points, which represents about 15% of the baseline turnover. Furthermore, while the effect on teachers moving to another school is statistically significant across our different specifications (around 25% of its baseline) the effect on the likelihood that teachers will leave the school system is only significant in some of our estimations. This suggests that the nature of these two decisions is different, and that the school labeling could be negatively affecting teachers' motivation to a point in which most of them are willing to leave a low-performing school only when they have another option to teach within the school system

The increased teacher mobility caused by the program was larger for certain groups of teachers. Specifically, less experienced teachers and those scoring lower on the college admission tests have a higher likelihood of leaving *in-recovery* schools. Given that we only have data on college admission outcomes for younger teachers, our results suggest that among less experienced teachers, the quality of teachers might have improved at *in-recovery* schools. However, this is not the case for more experienced teachers. Among those that have participated in the national teacher assessment, we fail to find a differentiated effect on the likelihood of teacher departure between teachers in the higher and lower categories of performance.

The combination of the findings of an increased likelihood of moving to other schools instead of moving out of the system among less experienced teachers, and the inability of *in-recovery* schools to attract new teachers, may be interpreted as indicating that the higher teacher mobility is related to both teachers avoiding *in-recovery* schools and school principals not making an effort to retain less experienced teachers, who are less likely to be effective (Clotfelter, Ladd, & Vigdor, 2006; Kane, Rockoff, & Staiger, 2008). Two other pieces of information seem to support this view. Elacqua et al. (2016) present suggestive evidence that schools in Santiago, the capital of Chile, initially reacted to the accountability system by implementing a series of practices aimed at improving student achievement in the short run, such as reallocating the most experienced teachers to fourth grade, the level where they measure the outcomes for the school classification, investing

less time and resources teacher training and evaluation. On the one hand, these practices might end up deteriorating the school's working conditions for teachers, especially for less experienced teachers seeking to learn during the first years of their careers. On the other hand, schools facing the urgency to improve their academic outcomes might also benefit from having fewer teachers that need training.

The overall effect is not clear, and more research is needed to assess how the school management and teacher responses to school accountability can ultimately lead to an improved school culture and higher academic outcomes. Regarding school management, it is important to determine if governments can introduce incentives to attract effective teachers, reallocate school resources more efficiently, and promote a different culture favoring students' engagement in the learning process. The way different groups of teachers perceive working at a school labeled as low performing is also informative to policy. It is possible that more and less effective teachers follow different approaches regarding their school classification, and determining which group considers it as a positive experience in their careers or as an experience that may hurt their future job prospects, can help improve the design of accountability systems.

While more research is still needed to determine whether the accountability system in Chile is benefiting or hurting low-performing schools, the failure of *in-recovery* schools to attract effective teachers to offset the turnover generated by the reform suggests that it may be an insufficient policy to raise student learning. Complimentary policies such as increasing monetary incentives to work in low-performing schools or targeting more resources and technical and pedagogical support to *in-recovery* schools could strengthen the potential effectiveness of this type of accountability reform.

References

Agencia de Calidad de la Educación de Chile (2019). Bases de datos. Santiago, Chile.

Aguirre, J. (2017). Can Progressive Vouchers Help the Poor Benefit from School Choice? Evidence from the Chilean Voucher System. Evidence from the Chilean Voucher System. *Unpublished manuscript (available at SSRN: https://ssrn.com/abstract=3123670)*.

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

Calonico, S., Cattaneo, M. D., & Titiunik, R. (2015). Optimal data-driven regression discontinuity plots. *Journal of the American Statistical Association, 110*(512), 1753-1769.

Calonico, S., Cattaneo, M., Farrell, M., & Titiunik, R. (2017). rdrobust: Software for regression discontinuity designs. *Stata Journal, 17*(2), 372-404.

Cattaneo, M. D., Jansson, M., & Ma, X. (2019). Simple local polynomial density estimators. *Journal of the American Statistical Association*, 1-11.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2006). Teacher-Student matching and the assessment of teacher effectiveness. *The Journal of Human Resources, 41*(4), 778–820.

Clotfelter, C. T., Ladd, H. F., Vigdor, J. L., & Diaz, R. A. (2004). Do school accountability systems make it more difficult for low-performing schools to attract and retain high-quality teachers? *Journal of Policy Analysis and Management, 23*(2), 251-271.

Dizon-Ross, R. (2017). How Does School Accountability Affect Teachers? Evidence from New York City. *Unpublished Manuscript*.

Elacqua, G., Martínez, M., Santos, H., & Urbina, D. (2016). Short-run effects of accountability pressures on teacher policies and practices in the voucher system in Santiago, Chile. *School Effectiveness and School Improvement, 27*(3), 385-405.

Elacqua, G., Santos, H., Urbina, D., & Martínez, M. (2011). *¿Estamos preparados para cerrar las malas escuelas en Chile? Impacto sobre equidad en el acceso a educación de calidad.* Santiago, Chile: Proyecto FONIDE 511083.

Feigenberg, B., Rivkin, S., & Yan, R. (2017). Illusory Gains from Chile's Targeted School Voucher Experiment. *National Bureau of Economic Research*, No. w23178.

Feng, L., Figlio, D. N., & Sass, T. (2010). School accountability and teacher mobility . *National Bureau of Economic Research*, No. w16070.

Feng, L., Figlio, D., & Sass, T. (2018). School accountability and teacher mobility. *Journal of Urban Economics, 103*, 1-17.

Friedman, M. (1955). The Role of Government in Education. En R. Solo, *Economics and the public interest* (págs. 123-144). New Brunswick, NJ: Rutgers University Press.

Gazmuri, A. (2015). School segregation in the presence of student sorting and cream-skimming: Evidence from a school voucher reform. *Unpublished Job Market Paper*.

Gill, B., Lockwood, J., Martorell, F., Messan, C., Booker, Vernez, G., . . . Garet, M. (2009). *An Exploratory Analysis of Adequate Yearly Progress, Identification for Improvement, and Student Achievement in Two States and Three Cities.* Washington, D.C.: U.S. Department of Education.

Gjefsen, H. M., & Gunnes, T. (2016). The effects of School Accountability on Teacher Mobility and Teacher Sorting. *MPRA Paper No. 69664*.

Goldhaber, D., & Hannaway, J. (2004). Accountability with a kicker: Observations on the Florida A+ accountability plan. *Phi Delta Kappan, 85*(8), 598-605.

Hsieh, C. T., & Urquiola, M. (2006). The effects of generalized school choice on achievement and stratification: Evidence from Chile's voucher program. *Journal of public Economics, 90*(8-9), 1477-1503.

Johnson, S. M., Berg, J. H., & Donaldson, M. L. (2005). *Who stays in teaching and why?: A review of the literature on teacher retention.* Project on the Next Generation of Teachers, Harvard Graduate School of Education.

Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2008). What does certification tell us about teacher effectiveness? Evidence from New York City. *Economics of Education Review, 27*(6), 615–631.

Ladd, H. F., & Zelli, A. (2002). School-based accountability in North Carolina: The responses of school principals. *Educational Administration Quarterly, 38*(4), 494-529.

Lee, D. S., & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature, 48*, 281-355.

Ley de Subvención Escolar Preferencial. (2008). *Ley N 20.248*. Biblioteca del Congreso Nacional de Chile. Obtenido de https://www.leychile.cl/Navegar?idNorma=269001

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of econometrics, 142*(2), 698-714.

Murnane, R. J., Waldman, M. R., Willett, J. B., Bos, M. S., & Vegas, E. (2017). The consequences of educational voucher reform in Chile. *National Bureau of Economic Research*, No. w23550.

Navarro-Palau, P. (2017). Effects of differentiated school vouchers: Evidence from a policy change and date of birth cutoffs. *Economics of Education Review, 58*, 86-107.

Neilson, C. (2013). Targeted vouchers, competition among schools, and the academic achievement of poor students. *Unpublished job market paper (Revise and Resubmit in Econometrica)*.

O'Malley, M. P., & Nelson, S. (2013). The public pedagogy of student activists in Chile: What have we learned from the penguins' revolution? *Journal of Curriculum Theorizing, 29*(2).

Papay, J. P., Willett, J. B., & Murnane, R. J. (2011). Extending the regression-discontinuity approach to multiple assignment variables. *Journal of Econometrics, 161*, 203-207.

Paredes, R., Bogolasky, F., Cabezas, V., Rivero, R., & Zahri, M. (2013). *Los determinantes del primer trabajo para profesores de Educación Básica en la Región Metropolitana. FONIDE F611105.* Santiago, Chile: Ministry of Education.

Ravitch, D. (2010). *The life and death of the great American school system: How testing and choice are undermining education.* New York, NY: Basic Books.

Reardon, S., & Robinson, J. (2012). Regression discontinuity designs with multiple rating-score variables. *Journal of Research on Educational Effectiveness, 5*, 83-104.

Reardon, S., Arshan, N., Atteberry, A., & Kurlaender, M. (2010). Effects of failing a high school exit exam on course. *Educational Evaluation and Policy Analysis, 32*, 498-520.

Robinson, J. (2011). Evaluating criteria for English learner reclassification: A causal-effects approach using a binding-score regression discontinuity design with instrumental variables. *Educational Evaluation and Policy Analysis, 33*(3), 267-292.

Robinson, J. P. (2011). Evaluating criteria for English learner reclassification: A causal-effects approach using a binding-score regression discontinuity design with instrumental variables. *Educational Evaluation and Policy Analysis*, 267-292.

Ronfeldt, M., Loeb, S., & Wyckoff, J. (2013). How teacher turnover harms student achievement. *American Educational Research Journal, 50*(1), 4-36.

Rouse, C. E., Hannaway, J., Goldhaber, D., & Figlio, D. (2013). Feeling the Florida heat? How low-performing schools respond to voucher and accountability pressure. *American Economic Journal: Economic Policy, 5*(2), 251-81.

Rouse, C. E., Hannaway, J., Goldhaber, D., & Figlio, D. (2013). Feeling the Florida heat? How low-performing schools respond to voucher and accountability pressure. *American Economic Journal: Economic Policy, 5*(2), 251-281.

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American psychologist, 55*(1), 68-78.

San Martín, E., Rivero, R., Bascopé, M., & Hurtado, C. (2013). *¿Es la prueba INICIA una medida predictiva de efectividad docente?* Santiago, Chile: Fondo de Investigación y Desarrollo en Educación.

Sims, S. (2016). *High-Stakes Accountability and Teacher Turnover: how do different school inspection judgements affect teachers' decisions to leave their school?* Unpublished manuscript, Department of Social Science, University College of London, England.

Van der Klaauw, W. (2008). Regression–discontinuity analysis: a survey of recent developments in economics. *Labour, 22*(2), 219-245.

Wong, V., Steiner, P., & Cook, T. (2013). Analyzing regression-discontinuity designs with multiple assignment variables: A comparative study of four methods. *Journal of Educational and Behavioral Statistics, 38*, 107-141.

Xu, K. L. (2017). Regression discontinuity with categorical outcomes. *Journal of Econometrics, 201*(1), 1-18.

Appendix 1: Results for main outcomes related to teacher mobility based on a categorical variable

To complement the estimations based on binary outcomes, we present the results of teacher mobility based on the work of Xu (2017) and the use of a categorical outcome variable with three categories defined according to teacher mobility the year following the school classification: i) leaves the school system; ii) leaves the school but not the school system; iii) stays in the same school. The main advantage of this approach is that the estimation incorporates the potential correlation among categories. The main disadvantage is that it does not support the computation of clustered standard errors or the inclusion of covariates.

The main results following this strategy are displayed in Table A1. Column (1) shows the effects of the program and column (2) the estimations associated with pre-treatment teacher mobility that serves as a placebo test. The overall result is consistent with the one presented earlier. Teachers are reacting to school accountability by leaving the *in-recovery* school the following year. The placebo test also shows that there is no effect of the program on pretreatment teacher mobility.

However, there are two differences compared to the results following the former strategy. First, the magnitude of the effect on the likelihood of leaving the school by moving to another is reduced from around 4 to 2.4 percentage points. This effect is still statistically significant at the 1% level of confidence. Second, the magnitude of the effect on leaving the school system does not change much (around 1.5 percentage points) but is statistically significant at the 5% level of confidence.

Table A1: RD estimates of the effect of working at an *in-recovery* school on teacher mobility using a categorical outcome variable (base category = Stays in the same school)

	Base category: Stays in the same school	
	2012-2015	2008-2011
	(1)	(2)
Number of schools	301-696	342-800
Number of teachers	6,771-16,337	7,672-17,868
Bandwidth	0.302	0.342
PANEL A: τ_1 - Leaves the system		
ATE	0.014**	-0.001
95% CI	[0.003, 0.033]	[-0.010, 0.018]
Robust 95% CI	[0.002, 0.035]	[-0.012, 0.020]
t-test; p-value	2.360; 0.018	0.567; 0.571
Robust t-test; p-value	2.161; 0.031	0.496; 0.620
PANEL B: τ_2 - Leaves the school		
ATE	0.024***	0.004
95% CI	[0.010, 0.045]	[-0.014, 0.018]
Robust 95% CI	[0.008, 0.047]	[-0.016, 0.020]
t-test; p-value	3.055; 0.002	0.214; 0.830
Robust t-test; p-value	2.795; 0.005	0.187; 0.852

∗ significant at 10%; ∗∗ significant at 5%; ∗∗∗ significant at 1%.

We also perform a robustness check replicating our main estimation with other six bandwidths. Three of these are bigger than the preferred bandwidth (1.75; 1.50; and 1.25 times) and three of them smaller (0.75; 0.50; and 0.25 times). In this case, our preferred bandwidth is the one suggested in Xu (2017) based on multinomial models. Table A2 presents the results of this exercise. The effect of the program on the outcome of leaving the school system is only statistically significant for bandwidths equal or greater than the optimal bandwidth. For the effect on the likelihood of leaving the school by moving to another, the result seems to be more robust. In only one of the six specifications is the effect not statistically significant at the 10% level.

Lastly, we estimate the effect of the program for the same six groups of teachers that we analyzed earlier. The new set of results are similar to those already presented. Less experienced teachers and those with a lower score on the college admission tests are more likely to leave the *in-recovery* school the year after the SEP classification. The magnitudes of these effects are similar to what we reported in our earlier analysis. Regarding the outcome of teachers leaving the school system, the magnitude of coefficients is close to that already reported but is now statistically significant for more experienced teachers (1.3 percentage points) and those with higher scores on the college entry exam (7.3 percentage points).

Table A2: Robustness of the teacher mobility estimates to the size of the bandwidth using a categorical outcome variable

	Base category: Stays in the same school					
	Bw = 0.25 · Opt BW	Bw = 0.50 · Opt BW	Bw = 0.75 · Opt BW	Bw = 1.25 · Opt BW	Bw = 1.50 · Opt BW	Bw = 1.75 · Opt BW
Number of schools	90-157	175-314	239-503	357-932	392-1,198	422-1,535
Number of teachers	2,096-3,691	4,011-7,316	5,413-11,808	8,058-22,038	8,869-28,739	9,427-37,129
Bandwidth	0.076	0.151	0.227	0.378	0.453	0.529
PANEL A: τ_1 - Leaves the system						
ATE	0.011	-0.004	0.010	0.010**	0.009**	0.006**
95% CI	[-0.017, 0.04]	[-0.025, 0.018]	[-0.005, 0.030]	[0.004, 0.031]	[0.007, 0.032]	[0.009, 0.032]
Robust 95% CI	[-0.017, 0.04]	[-0.025, 0.018]	[-0.006, 0.030]	[0.001, 0.034]	[0.002, 0.036]	[0.002, 0.039]
t-test; p-value	0.770; 0.441	-0.297; 0.766	-1.362; 0.173	-2.488; 0.013	-2.998; 0.003	-3.437; 0.001
Robust t-test; p-value	0.769; 0.442	-0.295; 0.768	-1.317; 0.188	-2.078; 0.038	-2.191; 0.028	-2.126; 0.033
PANEL B: τ_2 - Leaves the school						
ATE	0.063***	0.034***	0.022**	0.022***	0.01*	0.001
95% CI	[0.029, 0.098]	[0.011, 0.059]	[0.004, 0.045]	[0.012, 0.044]	[0.005, 0.0340]	[-0.001, 0.026]
Robust 95% CI	[0.029, 0.098]	[0.010, 0.059]	[0.004, 0.045]	[0.009, 0.047]	[-0.001, 0.0390]	[-0.009, 0.034]
t-test; p-value	3.615; 0.000	2.811; 0.005	2.384; 0.017	3.401; 0.001	2.564; 0.010	1.794; 0.073
Robust t-test; p-value	3.610; 0.000	2.784; 0.005	2.303; 0.021	2.835; 0.005	1.877; 0.061	1.118; 0.263

∗ significant at 10%; ∗∗ significant at 5%; ∗∗∗ significant at 1%.

Table A3: RD estimates of the effect of working at an *in-recovery* school on teacher mobility for different groups of teachers using a categorical outcome variable

	Base category: Stays in the same school					
	Unexperienced Teachers (<= 3 years)	Experienced Teachers (> 3 years)	Lower Teacher Assessment	Higher Teacher Assessment	Lower Score on College Admission Test	Higher Score on College Admission Test
Number of schools	299-721	343-861	333-1,465	340-1,333	619-1,695	672-2,591
Number of teachers	1,630-4,236	6,053-15,354	1,594-6,570	3,474-15,676	260-626	280-945
Bandwidth	0.323	0.358	0.602	0.528	0.331	0.466
PANEL A: τ_1 - Leaves the system						
ATE	0.029	0.013**	-0.005	-0.004	0.015	0.073***
95% CI	[-0.004, 0.066]	[0.003, 0.034]	[-0.028, 0.028]	[-0.017, 0.010]	[-0.035, 0.061]	[0.035, 0.134]
Robust 95% CI	[-0.007, 0.069]	[0.001, 0.036]	[-0.039, 0.039]	[-0.024, 0.017]	[-0.038, 0.064]	[0.026, 0.143]
t-test; p-value	1.728; 0.084	2.35; 0.019	0.005; 0.996	0.498; 0.618	-0.525; 0.600	-3.317; 0.001
Robust t-test; p-value	1.600; 0.110	2.032; 0.042	0.004; 0.997	0.331; 0.741	-0.491; 0.623	-2.837; 0.005
PANEL B: τ_2 - Leaves the school						
ATE	0.090***	0.010	0.019	-0.024	0.138***	-0.015
95% CI	[0.059, 0.149]	[-0.006, 0.028]	[-0.003, 0.061]	[-0.038, 0]	[0.092, 0.227]	[-0.063, 0.066]
Robust 95% CI	[0.055, 0.153]	[-0.009, 0.03]	[-0.017, 0.074]	[-0.047, 0.009]	[0.087, 0.232]	[-0.071, 0.074]
t-test; p-value	4.536; 0.000	1.248; 0.212	1.763; 0.078	-1.927; 0.054	4.659; 0.000	0.047; 0.962
Robust t-test; p-value	4.192; 0.000	1.082; 0.279	1.243; 0.214	-1.305; 0.192	4.319; 0.000	0.042; 0.967

∗ significant at 10%; ∗∗ significant at 5%; ∗∗∗ significant at 1%.

Considering all results, the most consistent is that teachers working at *in-recovery* schools are reacting to the SEP classification by moving to another school the following year. It also appears that less experienced teachers and those that performed worse on the college entry examination are the most likely to move. A less robust result, which appears when we follow the estimation strategy proposed by Xu (2017), is that the accountability program in Chile might also have a weaker effect on a group of teachers that decided to leave the profession, especially among those that are more experienced and performed better in the college admission tests.

Appendix 2: School accountability and student achievement: An alternative definition of control group

In this appendix we replicate results from section 4.4 but redefining the control group of schools. Specifically, we exclude schools that were not classified as in-recovery in 2012 and classified as *in-recovery* at least once between 2013 and 2015. We perform this exercise as a robustness check to ensure that after excluding schools from the control group that were also treated at some point the results are maintained.

Table A4 shows the results of the estimations using this new control group. The results presented in the main body of the paper hold. The school accountability system in Chile seems not to have improved student achievement.

Table A4: RD estimates of the effect of being classified as *in-recovery* on student achievement on math

	SIMCE Math						
	2011	2012	2013	2014	2015	2016	Value added Baseline 4th grade 2011
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
4th grade							
Treatment effect	-0.25	0.046	-0.206	-0.032	-0.165	-0.374	-
	(0.2)	(0.239)	(0.229)	(0.266)	(0.236)	(0.241)	-
Bandwidth	0.400	0.380	0.390	0.350	0.360	0.350	-
Number of Obs: Left - Right	134-256	128-230	132-235	114-205	118-212	110-207	-
6th grade							
Treatment effect	-	-	-0.157	-0.256	-0.134	-0.092	0.037
	-	-	(0.12)	(0.165)	(0.182)	(0.18)	(0.248)
Bandwidth	-	-	0.500	0.350	0.390	0.350	0.340
Number of Obs: Left - Right	-	-	151-347	114-199	129-233	110-199	99-171
8th grade							
Treatment effect	-0.169	-	0.045	-0.102	-0.204	-	0.049
	(0.159)	-	(0.157)	(0.116)	(0.182)	-	(0.295)
Bandwidth	0.34	-	0.39	0.55	0.37	-	0.3
Number of Obs: Left - Right	113-192	-	131-235	145-398	122-218	-	74-124

Standard error were calculated using clusters at the school-year level and are shown in parenthesis