Short-run effects of accountability pressures on teacher policies and practices in the Chilean voucher system

Gregory Elacqua
Matias Martinez
Humberto Santos
Daniela Uribe
Public Policy Institute
Universidad Diego Portales

This research analyzes the impact of the Adjusted Voucher Law´s school rankings on low-performing schools in Chile and provides evidence on the effects of the pressures of accountability systems on teacher policies and practices. The empirical strategy is based on the fact that schools are ranked according to their position on a set of thresholds. We used a generalization of the traditional regression discontinuity design for the case where treatment assignment is determined by n variables. To gather information on teacher policies and practices, we conducted a survey of fourth-grade teachers in the Greater Santiago area. The results indicate that low-performing schools responded to the treatment by implementing policies that seek to improve their results in the short term. We also found no significant effects on teaching practices, suggesting that many of these changes are implemented top-down from the school administrators, without involving teachers in the process.

Keywords: school accountability, teacher policy, Chile

Introduction

Over the past 20 years one of the major trends in education reform in countries on every continent has been the implementation of high-stakes testing and school accountability. Even though several countries have been using standardized tests for measuring student achievement for several decades, the innovation of accountability reform is the use of student outcomes to evaluate teacher and school performance (Elmore, Abelman, & Fuhrman, 1996; Figlio & Loeb, 2011; O'Day, 2002). Under a school accountability system, the State sets performance standards and provides rewards or sanctions to schools that meet (or fail to meet) these standards. Schools are often ranked according to their performance and have a specific period of time to improve their outcomes. If they do not meet the standards on time, schools face sanctions that range from mandatory improvement plans to school closure.

The wide implementation of accountability systems in the United States, Europe and some countries in Latin America has generated a raging debate on the effects of these systems on school performance. Advocates argue these mechanisms have positive effects on academic outcomes in low-performing schools. However, critics maintain that these improvements are not explained by real progress in student learning, but rather by strategic behavior that schools develop as they internalize accountability pressures.

Research over the past decade has focused mainly on achievement gains, while studies on how these interventions actually modify schools' day-to-day practices and policies are scarce (Booher-Jennings, 2005). This paper contributes to this debate by analyzing the effects of accountability pressures on teacher's policies and practices in Santiago, Chile. This study is one of the first evaluations of Chile's school accountability system, so we expect that our results will improve our understanding about how schools responded to this policy. Furthermore, this article contributes to the scholarly debate on the effects of accountability pressure on school behavior, which has been mainly focused on the United States and Great Britain. The Chilean education system is an interesting case for comparison since it presents a high-stakes accountability scheme within a school choice institutional arrangement.

Chile is one of the few countries that, as part of a comprehensive neoliberal reform introduced by the military regime, implemented a universal voucher program that has provided school choice for parents since 1981. Under this scheme, school quality was supposed to be assured by parent accountability. Similar to what Hirschman (1970) proposed for companies, when schools offer a low-quality education, parents have two options: they can leave the school ("exit") or they can express their dissatisfaction ("voice"). In a competitive schooling market, choice advocates maintain that low-quality schools would disappear, because they will lose students as a result of the exit and voice mechanisms (Hirschman, 1970). However, in the mid-2000s, despite substantial increases in funding and parental choice, the performance of Chilean schools was still poor compared with OECD countries and national test scores were stagnant. In response to this scenario, in 2008 Congress enacted a new education reform that, among other changes, implemented a system of state accountability for schools. The *Subvención Escolar Preferencial* Law (Adjusted Voucher Law—SEP) created incentives for school leaders and teachers to improve student performance. Similar to other state accountability systems, SEP establishes minimum performance standards and ranks schools according to their performance on a national standardized test and other outcomes. It also establishes sanctions for low-performing schools, including closure when a school does not show adequate improvement.

This article is organized as follows.The first section lays out the theoretical framework and reviews the empirical literature on the effects of accountability on different outcomes. The second section describes the school choice and accountability system in Chile. The next two sections discuss the methodology we employ and our data. The fifth section describes our results and the final section concludes and discusses some policy implications and directions for future research.

Theoretical framework and literature

Accountability mechanisms have been implemented in various educational systems around the world. Examples include the No Child Left Behind Law (2001) in the United States, the Education and Inspection Law (2006) in England, and the SEP Law (2008) in Chile. In all of these examples, the State instituted accountability systems, establishing performance goals in schools and sanctions for those that do not meet them. The identification, classification, and subsequent publication of low performing schools are all key components of the three accountability systems (Figlio & Lucas, 2004). The objective of these actions is to increase the supervision of low-performing schools by parents and the government and to increase the pressure on schools for improvement (Jacob, 2005). Previous research shows that the mere identification of low-performing schools operates as a social stigma for its principals, teachers, and students, increasing pressure to improve (Goldhaber & Hannaway, 2004; Ladd & Glennie, 2001). Once the low-performing schools are identified, different measures are gradually implemented to reverse the situation, with the ultimate consequence of closing the school (Brady, 2003; Hanushek & Raymond, 2003; Spreng, 2005). The assumption is that closing chronically deficient schools would operate as an incentive for other low-performing schools to improve under the threat of closure (Smarick, 2010).

Among the few studies that have analyzed the effect of accountability pressures on teaching policies and practices, Rouse et al. (2013) show that schools under high accountability pressures in Florida modified some of their internal practices and policies, especially those referring to teachers. The authors maintain that these changes explain performance gains in these schools. This is consistent with the results of other studies in the United States, which demonstrate that after the implementation of accountability mechanisms in New York, Chicago, and Texas, low-performing schools improved their performance (Deming, Cohodes, Jennings, & Jencks, 2013; Figlio & Rouse, 2006; Jacob, 2005; Rockoff & Turner, 2010). Similarly, studies in the United States show that accountability systems explain test score

improvements at the national level (Carnoy & Loeb, 2002; Hanushek & Raymond, 2005; West & Peterson, 2006).

Critics have argued that accountability pressures also produce undesirable effects once schools internalize them. First, given that the performance standards set by the State measure only certain subjects from the curriculum, researchers have documented that schools dedicate more time to courses that are evaluated in comparison with those that are not (Hannaway & Hamilton, 2007). For example, in Kentucky, where students are evaluated in fifth grade, 82% of 5th-grade teachers stated that they increased the instruction time for math, in contrast with 14% of fourth-grade teachers. (Stecher & Barron, 2001). Similar results were found in Washington, California, Georgia, and Pennsylvania (Hamilton et al., 2007; Stecher et al.,2000). Although part of the evidence indicates that the origin of these changes is in the teachers themselves, who react to accountability pressures, it has also been demonstrated that principals and local authorities encourage these practices. In Florida, Hannaway and Cohodes (2007) show that after the implementation of accountability measures, schools modified their policies on the instruction time of certain courses, increasing the hours for subjects evaluated on the standardized tests. In North Carolina, Ladd and Zelli (2002) find that many principals encourage teachers to spend more time on language and math, as well as reallocating resources from other subjects to strengthen these two areas.

Another effect of accountability pressures has to do with the abilities that teachers emphasize in their classrooms. There is evidence that teachers modify their teaching strategies based on the instrument with which their students are evaluated (Elliot & Borko, 1999; Wolf, Borko, McIver, & Elliott, 1999). For example, in Vermont, the evaluation tool corresponds to the development of a math portfolio. In response, teachers emphasized the teaching of skills for problem solving (Koretz, Stecher, Klein, & McCaffrey, 1994). In the case of standardized tests, teachers respond to these evaluations by dedicating part of their teaching to basic skills, vocabulary, and multiple-choice questions (Romberg, Zarinnia, & Williams, 1989; Shepard & Dougherty, 1991). Moreover, it has been demonstrated that teachers respond to this type of evaluation by using less frequently other strategies, such as essays, written responses, and other activities not evaluated on standardized tests (Shepard & Dougherty, 1991; Smith & Rottenberg, 1991). The study by Pedulla et al. (2003) in the United States concluded that the effect of focusing classes on abilities tested by evaluation instruments is greater in states where the consequences of accountability are more severe.

Accountability pressures have also led teachers to try to "outsmart" standardized tests through various practices. First, some teachers alter the pool of students evaluated. For example, Cullen and Reback (2006), Figlio and Getzler (2006), and Jacob (2005) find that some teachers reclassify their low performing students as students with learning disabilities so that their scores are not counted. Figlio (2006) finds that some schools suspend low-performing students the day of the test. Jacob and Levitt (2003) find that under accountability systems, teachers have a greater probability of helping students answer the tests. For example, in Chicago, in 4% of the classes teachers changed or filled in student answers (Jacob & Levitt, 2003). Pedulla et al. (2003) find that teachers provide 12% to 19% more time than stipulated for students to take the tests. There is also evidence that teachers often give more attention to students who are closer to surpassing the performance threshold established by the authorities, disregarding students who are far below or above the threshold (Booher-Jennings, 2005; Hamilton et al., 2007; Neal & Schanzenbach, 2007).

In short, there is a persistent debate about the effects of school accountability on teaching policies and practices, beyond standardized test results. The empirical evidence available to date is mixed, which emphasizes the need to study the impact of this type of policy under different accountability and choice designs.

School accountability in Chile: the Subvención Escolar Preferencial Law (Ley SEP)

The military regime introduced a sweeping education reform package in 1981. First, they decentralized the administration of public schools from the central level (the Ministry of Education) to the local level (municipalities). Second, they altered the funding scheme by tying public school funding to the number of students enrolled. After these changes were introduced, the Chilean educational system was composed of three types of schools, depending on their type of administration and source of funding: *municipal* (public) schools, which are financed with government subsidies and administered by the local municipal government, whose maximum authority is the mayor; *private voucher* schools, also financed with government subsidies, but administered by a private (for-profit or nonprofit, religious or secular) organization, and finally *private* schools, which are financed and administered privately. Although the system's basic structure remained in placefor almost three decades, in 2008 an important modification was introduced with the enactment of the Subvención Escolar Preferencial Law (Adjusted Voucher Law—SEP).

The SEP Law is the first initiative that introduced explicit school accountability mechanisms in Chile. In simple terms, the law introduced an additional voucher (about 50% over the base voucher) for students classified as vulnerable (*priority* students[1]), who attend municipal or private voucher schools, and who voluntarily agree to participate in the program,[2] under the condition of meeting minimum standards for academic performance. In 2012, SEP gave an additional voucher to schools for each of their *priority* students between pre-kindergarten and eighth grade. For example, an elementary school with a total enrollment of 300 students, with half of them classified as *priority* students, received an additional subsidy of 41% (about USD$14,500 per month).

In order for schools to receive the additional voucher, they must also meet a series of requirements[3] and must sign an agreement with the Ministry of Education, which includes a series of commitments: documenting the use of resources, establishing effectiveness goals for students' academic performance, and providing parents with information on school performance. One of the most important requirements is that all schools participating in the SEP program must develop and carry out an improvement plan (Plan de Mejoramiento Educativo—PME), led by the school principal with the participation of the rest of the school community. The PME requires actions in four areas: i) curriculum management; ii) school leadership; iii) school climate; and iv) managing school resources. Schools have the option of hiring technical educational assistance to develop their PME, provided directly by the Ministry of Education or by registered external agencies. These agencies provide consulting, training, evaluation, and institutional diagnostic services.

Chile´s accountability program ranks schools into three categories: i) *autonomous* (schools that systematically perform above national standards); *emerging* (schools that do not systematically perform above national standards); and *recovering* (schools that systematically perform below national standards). The ranking has consequences for low-performing schools.

Schools were not ranked in the recovery category during the first four years of the SEP Law. 2012 was the first year that schools were classified in this category.[4] This change in the program allows researchers to evaluate the effects of the threat of accountability pressures on schools that do not meet minimum performance standards.

Specifically, the main objective of this research is to estimate the causal impact of being classified as a recovering school on teaching policies and practices. There are two components in the SEP Law that introduce specific threats to low-performing schools. First, and unlike autonomous and emerging schools, recovering schools have a tighter deadline to improve their results. If the school does not manage to move to a higher category in three years, the Ministry of Education will report this to the school community and will encourage families to consider another schooling option for their children, as well as facilitating transportation to a better school. However, if the school remains in the recovering category for four years, the Ministry will revoke its license to operate and receive public funding.[5]

Second, information on the school ranking will be widely disseminated among families, which is intended to influence parental preferences in school choice. Thus, being classified as low-performing could have a negative effect on future enrollments and on the characteristics of families (e.g. motivation) who are willing to choose such a school.

We hypothesize that recovering schools will respond to the threats of accountability by adopting policies and practices that aim to improve their students' performance. The channels through which such pressure could generate change are diverse and involve the entire school community. First, the most direct channel is the threat faced by teachers and principals of losing their jobs as a consequence of enrollment reduction and eventual closure, if the minimum standards are not met by the stipulated deadlines. In the case of the United States, for example, Reback, Rockoff, and Schwartz (2011) find that teachers working in schools that face greater pressures to improve in the short-run have a greater sense of job insecurity. Second, given that the school ranking will be available to parents, being classified as recovering can be in and of itself an important "social stigma" for schools. Indeed, some researchers have shown that being a teacher at a low-performing school implies a stigma and a loss of well-being (Ravitch, 2010; Rouse et al., 2013). According to some authors, being identified as a teacher or principal whose school cannot provide a high-quality education to its students implies a loss status in their community (Rouse, et al., 2013). There is also evidence that parents respond to their school's ranking, especially when it is classified as low-performing, which could influence their relationship with the principal and teachers (e.g. Rockoff & Turner, 2010). Finally, some authors argue that the pressure to improve performance leads teachers to raise their expectations of the potential of low-performing students, which leads them to change their traditional teaching practices in the classroom (Koschoreck, 2001; Scheurich, Skrla, & Johnson, 2000).

This research evaluates the marginal impact of a specific component of the SEP classification that affects exclusively recovering schools. Other studies, mainly in the U.S., have also evaluated this marginal "threat effect" (e.g. Chakrabarti, 2008; Gill et al., 2009; Greene & Winters, 2003; Rouse et al., 2013; West & Peterson, 2005;). In order to measure global effects, we would need to analyze changes over time or compare schools who participate in SEP with those who do not, which is beyond the scope of this study. There may be global effects that impact all schools, regardless of their classification. For example, the threat of being ranked in the recovering category can also lead emerging schools to change their teaching policies and practices. In this sense, no school can consider itself totally unaffected by the treatment.

One of the challenges of the present study is that very little time has passed since this policy was instituted. If there is inertia in schools' behavior, perhaps the expected changes will occur over a longer period of time. However, there is evidence that schools respond in the short-term to threats of sanctions introduced by accountability systems. For example, Rockoff and Turner (2010) find that accountability pressures in New York had positive effects on standardized test results and on parent satisfaction four to six months after schools were ranked. The authors also show that the impact was greater in schools ranked as low-performing.

Methodology

In order to estimate the causal effect of being classified as a recovering school, we exploit the fact that the methodology used to rank schools is based on a school's position relative to a set of variables and their respective thresholds. These variables include national standardized test scores, the number of students tested, the number of available measurements, and a set of indicators that measure other quality dimensions (e.g. pass rate and teacher evaluation results). Unlike a traditional regression-discontinuity design (RDD),[6] where units are assigned to treatment and control conditions based on a

single cutoff score on a continuous assignment variable, in this case we used a generalization for the case where multiple assignment variables and cutoffs may be used for treatment assignment.

Multivariate regression-discontinuity design

In many cases, treatment assignment depends on the value of a set of variables. However, an RDD with multiple assignment variables (multivariate regression-discontinuity design or MRDD) raises challenges that are distinct from those identified in a traditional design, because analytic 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 the assignment variable. These challenges have recently been addressed in a series of studies (Imbens & Zajonc, 2011; Papay, Willet, & Murnane, 2011; Reardon & Robinson, 2012; Wong, Steiner, & Cook, 2013).

Reardon and Robinson (2012) show that different methods to estimate average treatment effects with multiple assignment variables are based on regression models as follows:

$$Y_i = m(R1_i, R2_i, \dots, Rn_i) + \sum_k \tau_k T_i^k + X_i B + e_i, \quad (1)$$

where $\{R1_i, R2_i, \dots, RJ_i\} \in \boldsymbol{D} \subset \boldsymbol{R}$.

$R1_i, R2_i, \dots, Rn_i$ correspond to the n assignment variables and T_i^k is a dummy variable indicating if unit i is assigned to treatment k. The estimators of treatment effects τ_k differ in two important ways: i) the specification of the m function and ii) the $\boldsymbol{D}$ domain of observations used in estimating the model, which is a subset of the space formed by the n assignment variables $(\boldsymbol{R})$. The inclusion of pretreatment covariates (X_i) may increase the precision of the estimates, but is generally unnecessary, as the model is well identified without it (Imbens & Lemieux, 2008; Lee & Lemieux, 2010). The choice of the functional form of m may be important, especially when there are few observations near the frontier. In this case, it is necessary to use data further from the cutoff score and assumptions about the functional form of the average potential outcome surfaces, but doing so increases the potential bias in the estimation. In other words, there is a trade-off between bias and precision.

Reardon and Robinson (2012) present five estimation methods: *response surface* RD, *frontier* RD, *fuzzy frontier* RD, *distance-based* RD, and *binding-score* RD. In this paper, we use the binding-score method because it has advantages over other approaches when there are few observations in the dataset.[7]The main advantage of this approach is that it allows the researcher to collapse scores from multiple assignment rules into a single assignment variable and therefore can use all of the observations simultaneously in the estimation. The approach also generalizes well to MRDDs with more than two assignment variables and simplifies the analyses for estimating average treatment effects across multiple discontinuity frontiers. This method has been used, for example, in the evaluation of NCLB in the United States (e.g. Ahn & Vigdor, 2009; Gill et al., 2009). Others examples are found in Martorell (2005), Reardon, Arshan, Atteberry, and Kurlaender (2010), and Robinson (2011). One disadvantage is that it does not allow the estimation of frontier-specific effects, so we cannot test the existence of heterogeneous treatment effects.

Binding-Score RD

This method is based on the construction of a new assignment variable Z (*binding-score*) that perfectly determines treatment assignment. For example, suppose that treatment assignment depends

on two variables (let's call them R and M). In particular, schools are assigned to a single treatment condition T if they score below both cutoffs (let's call them 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:

$$Z_i = max(R_i^c, M_i^c), \tag{2}$$

where $R_i^c = R_i - r_c$ y $M_i^c = M_i - m_c$. By construction, $T_i = 1$ if $Z_i < 0$ and $T_i = 0$ if $Z_i \geq 0$[8]

Thus, the problem becomes a traditional RDD and then all of the standard analytic methods can be used, defining Z as the assignment variable and zero as the cutoff. Although this transformation applied to the original assignment variables, Wong, Steiner, and Cook (2013) show that this method estimates the same causal effect as alternative methods.

Data

The main source of information for the development of this research was the national SEP classification database for the year 2012. The database contains the school ranking and all of the assignment variables. The most important variable is the school's performance on the Chilean standardized test (Sistema de Medición de la Calidad de la Educación—SIMCE[9]), for fourth-grade students during the last three years for which information is available. Besides SIMCE results, the SEP classification incorporates a set of complementary indicators that measure other dimensions of educational quality. These indicators include the percentage of students who achieve the national standard, the percentage of students who remain in school until the end of the year, quality of working conditions, participation of teachers and families in the development of the school's educational project, the school's capacity to incorporate educational innovation, and the results on the national teacher evaluation (only for public schools).

This dataset was merged with the school panel, which includes information on student achievement, student demographics, and available schooling inputs for all Chilean schools since 1990. We restrict the data set in several ways. First, it includes only schools that participate in SEP. Second, the universe is restricted to urban schools in the Greater Santiago area. Third, autonomous schools are excluded. Therefore, the control group is composed only of emerging schools. The justification for this restriction is that emerging schools, similar to recovering schools, are required to present an improvement plan (PME), with a diagnosis of the school's initial situation and establish a set of goals for educational improvement. This allows us to isolate the effect of being classified as recovering from the required changes introduced by the SEP Law. Fourth, the school universe was limited to those having two or more SIMCE measurements. Finally, a group of schools ranked as emerging because they had fewer than two available measurements was also excluded. The total universe was composed of 87 (12.6%) recovering schools and 602 (87.4%) emerging schools. The binding-score (Z) was constructed from this final dataset. Details on the construction of this variable are presented in the appendix.

To gather information on teaching policies and practices, we conducted a survey with a sample of fourth-grade math teachers[10] from April to June 2012. The target sample size was obtained from the formulas presented in Schochet (2009), who derives the sample size necessary to identify causal effects in RDD. On the basis of this, we obtained a sample of 181 schools (87 recovering schools and 94 emerging schools), located within a window of ±0.71 units of Z_i around the threshold ($Z_i = 0$).[11]

Additionally, for each one of the emerging schools in the sample, three replacement schools were identified, which correspond to the three closest "neighbors" according to the value of Z_i.

The survey questionnaire was based on empirical literature about the effect of accountability pressures on teaching policies and practices and from a compilation of teacher surveys carried out in Chile. The questions were grouped into various domains: i) teacher incentives; ii) policies focused on low-performing students; iii) internal teacher evaluations; iv) policies focused on low-performing teachers; v) SIMCE preparation; vi) teaching strategies in the classroom; vii) organization of group work; viii) internal evaluations applied to students; ix) parent involvement; x) qualitative SEP evaluation; xi) teachers' educational background and working conditions; xii) teacher socioeconomic characteristics; and xiii) the distribution of weekly class hours among different subjects.

We surveyed 134 teachers,[12,13] 54 from recovering schools and 80 from emerging schools. Among the latter, 22 cases were replacements. Table 1 presents a summary of the sample's descriptive statistics, comparing teachers from recovering and emerging schools. Based on a difference in proportions test and a difference in means test, we conclude that there are few statistically significant differences. [14] The results indicate that teachers in recovering schools are less likely to be exposed to an internal teacher evaluation (38.5% vs. 55.1%); they are less likely to receive support for class preparation if they have low performance (40.7% vs. 57.0%); they are less likely to use multiple-choice tests to evaluate their students (90.7% vs. 97.5%); and they are less likely to arrange face-to-face meetings with parents (24.5% vs. 43%).On the other hand, emerging schools, as expected, enroll students of a higher socioeconomic level than recovering schools, but no significant differences are found in terms of available school resources (see Table 1).

Beyond the differences between teachers in these two types of schools, it is interesting to note that SIMCE has an important influence on classroom activities and evaluation methods used by teachers. Thus, for example, over 70% of teachers in recovering and emerging schools responded that the school principal had established minimum performance goals for the SIMCE test; around 60% said that they had used exercises similar to SIMCE in their classrooms and had trained students how to answer multiple-choice questions every day or almost every day; and over 70% said that they had used practice SIMCE tests to evaluate their students. Finally, on average, almost 60% of total weekly teaching hours are dedicated to subjects covered on the SIMCE test (i.e. language, math, and natural sciences).

On the other hand, we observe that over 90% of teachers say that they are familiar with the SEP law and know that it ranks schools mainly by SIMCE results. However, most of them do not know the category in which their school was classified. In fact, less than 30% of teachers surveyed (both in recovering and emerging schools) knew their school´s ranking.

[Table 1]

Results

Testing the validity of regression discontinuity design

Under the assumption that units cannot manipulate the assignment variable, the causal inference of an RDD is equivalent to that which would be obtained from a random experiment (Lee, 2008). As McCrary (2008) points out, a direct way to test whether there is a precise control of the assignment variable (prior to treatment), is to examine its density function. A discontinuity in this function would be evidence of some degree of "sorting" around the threshold. In the case of the SEP law, if schools could precisely manipulate the variables that determine its classification, many of them

would be located immediately after the thresholds of assignments variables—avoiding being classified as recovering—which should generate a discontinuity in the density function of these variables.

However, there are at least two features of the SEP classification that reduce the likelihood that schools can manipulate assignment variables. First, the ranking, along with its methodology, was officially reported to schools during the month of September 2011. Since the 2012 ranking depends on the SIMCE results of 2008, 2009, and 2010, schools had no certainty about the methodology that would be used at the time of the measurements considered for the classification. Second, the Ministry of Education conducts a monitoring process during and after the application of the SIMCE test by hiring external staff for schools, seeking to ensure that all fourth-grade students in the country take the tests. Teachers at each school can help the external examiner with the organization and discipline of the students, but they neither have access to the tests nor to the classrooms where they are applied.

McCrary (2008) proposes a two-step procedure to test if there is a discontinuity in the density of the assignment variable. In the first step, a histogram of this 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 (McCrary, 2008). Table 2 shows the test[15] value for the eight assignment variables used for the SEP classification and for the binding-score and Figure 1 illustrates the estimated density of Binding-score (Z_i). The results show that there is no evidence to reject the null hypothesis of continuity in the assignment variable densities at a confidence level of 95%. The only exception is the $p250_{i,2008}$variable. The above results provide evidence that schools do not have the ability to accurately manipulate the variables that determine their SEP classification, which gives validity to the estimates we present in the following section.

[Table 2]

[Figure 1]

According to Lee and Lemieux (2008), if the assumption that there is no precise manipulation or sorting of the assignment variable is valid, then there should be no discontinuities in variables that are determined prior to the assignment. We use MINEDUC administrative data to check for discontinuities in a number of pre treatment variables. No discontinuities were found between recovering and emerging schools on annual resources received by SEP, the socioeconomic composition of schools, and teaching practices.[16]

Estimation and results

Table 3 presents the results of estimates from a regression model with the following functional form:

$$Y = \alpha_l + \tau T + \beta_l(Z - z_c) + (\beta_r - \beta_l)T(Z - z_c) + \varepsilon. \quad (3)$$

Where Y is the analyzed result, T takes the value of one if the school is classified as recovering and zero for emerging, $(Z - z_c)$ represents the distance from the school to the threshold of the assignment variable constructed with the binding-score method and ε is an error term with normal distribution.

In order to analyze the robustness of the results, we performed three independent exercises. First, with the goal of capturing possible nonlinearities in $\mathrm{E}[\mathrm{Y_i}|\mathrm{Z_i} = \mathrm{z}]$ that may be confused with discontinuities, we incorporate higher order terms (polynomial regressions) of the assignment variable (Z). Second, previous models were reestimated using a smaller bandwidth around the threshold (± 0.32 Z_i units). Finally, a vector of pretreatment variables (X) was introduced into the regression. While it is true that the addition of these variables should not alter the RD identification strategy, many studies use control variables to eliminate potential bias in small samples, which may be relevant in our case because the estimates consider observations distant from $Z = z_c$. Additionally, the inclusion of vector X in the regression improves the accuracy of the estimates if these variables are correlated with potential results. In fact, this is a common practice even in analyses of randomized experiments. Table 4 shows the details of the pretreatment variables included in the regressions.

In total, we evaluated 86 outcome variables, consisting of school policies, teaching practices, and teacher characteristics. Tables 3 and 5 include only those outcome variables (Y) for which the average treatment effect is statistically significant in at least one of the eight specifications and that maintain the sign in all specifications. [17, 18]

Regarding school policies, we observe the application of two types of policies in response to the treatment: i) policies targeted for low-achieving students, and for both high and low-performing teachers, and ii) less use of internal accountability. In the former case, we specifically found that recovering schools are more likely to use a personal tutor after school to improve the performance of low-achieving students (16.2 to 33.8 percentage points (pp)), less likely to reward high-performing teachers by sponsoring them for training courses or seminars (-17.6 to -35.5 pp), and less likely to require low-performing teachers to attend training courses (-23.9 to -43.5 pp). Meanwhile, specific results for the second type of policy show that, for teachers working at recovering schools, there is a lower likelihood that their classes have been observed by an external specialist in the last four weeks (-19.5 to-63.4 pp) and there is a lower likelihood of having observed a class of one of their colleagues in the same period (-22.9 to -93.8 pp). Moreover, recovering schools are also less likely to have an internal teacher evaluation (-17.6 to -60.0 pp).

With regard to teaching practices, the results suggest modest changes in response to treatment. Specifically, we find that teachers at recovering schools are more likely to have evaluated students through project reports (7.3 to 46.0 pp) and to have used multiple-choice tests (10.8 to 36.8 pp) in the last four weeks.

Finally, we found discontinuities between these two types of schools in several teacher characteristics, which may be attributable to the SEP classification. In this dimension, the results indicate that teachers at recovering schools have employment contracts that stipulate more instruction hours (1.2 to 11.1 hours), they are more likely to have a graduate degree (6.6 to 73.2 p.p.), they are less likely to have a teaching degree without a specialization (-20.9 to -59.0 pp), they are more likely to have studied at an institution that had higher exit competency exam scores[19] (6.4 to 15.5 pp more in terms of correct answers), and they have more years of teaching experience (4.8 to 10.9 years).

As a summary, Figure 2 shows graphically four key findings of our research about the response of recovering schools to accountability pressures: a higher use of personal tutors after school, a lower mandate for attending teacher training courses, a lower percentage of teachers observing one of his/her colleagues conducting a class, and higher exit competency exam results (2010) from teacher´s undergraduate school of education. Each graph shows the averages for each outcome variable(Y) in bins of 0.1units of assignment variable (Z). Furthermore, we plot the linear model and a non-parametric approach by local regressions.

[Table 4]

[Table 5]

[Table 6]

[Figure 2]

Discussion

One of today's most controversial topics in education reform discussions is school accountability. The present study seeks to contribute to this debate by analyzing the effects of accountability threats on teacher policies and practices under the SEP Law in Chile.

Our results indicate that low-performing schools respond to the accountability pressures generated by the SEP Law. Specifically, our analysis shows two important effects. First, low-performing schools respond by implementing policies that seek to improve students' academic performance in the short-run. Second, the principal changes made in recovering schools were on the level of teaching policies and not practices; moreover, these changes were likely implemented top-down by the school principals without involving teachers in the process.

Regarding the first finding, we observe that low-performing schools, compared to their counterfactual, are less likely to implement teacher evaluation systems and fund teacher training courses, and are more likely to introduce after-school tutoring programs for low-performing students. The choice of these types of policies suggests that principals prefer to adopt the most time-efficient measures, those that are the most likely to increase student performance in the short term. It is important to note that these results do not mean that low-performing schools have decreased the amount of financial resources for teacher training or internal accountability. A more accurate interpretation is that, in the absence of the threat they face, these schools would have allocated a larger share of the additional resources provided by SEP to similar policies as emerging schools adopted.

We also find that recovering schools are more likely to assign their most qualified teachers to the fourth-grade, since that is the level where SIMCE evaluations are used for the SEP classification. In recovering schools, fourth-grade teachers are hired for more classroom hours, have more teaching experience, and a higher quality educational background. These results could be explained by two measures that recovering schools have adopted: i) hiring new teachers with these characteristics or ii) reassigning these teachers to the tested grade level. We posit that the second alternative is more likely since we did not find a discontinuity in average teacher seniority.

The adoption of short-term teaching policies in recovering schools could be a response to the conditions in which the SEP Law is carried out. According to the current methodology, the results obtained by students in recovering schools in 2012 (post treatment) will only be considered for the school classification, and in addition to the results from 2010 and 2011 (pretreatment), in the school year of 2014. If in that year's ranking, schools remain in the recovering category, the Ministry of Education will inform parents of other higher performing neighborhood alternatives, and alerting them of the possibility of the school closing the following year. This creates a strong incentive for principals to seek immediate results since any plans that consider longer periods may jeopardize the survival of the school.

Low-performing schools seem to be responding in a way that is consistent with the incentives and deadlines of the SEP design. This highlights the importance of the design of accountability systems,

because the deadlines, types of sanctions, grades to be evaluated, and communication strategy, among other elements, are key in determining how schools under threat will change their teaching policies and practices. In the particular case of SEP, it is relevant to ask if the changes found at recovering schools have desirable effects for the quality of education. Although this article does not analyze the effect of these policies on academic outcomes, our preliminary evidence on the assignment of the most qualified teachers to the fourth-grade is an effect that could be negative for students from other levels. That is, unless high-quality teachers were recruited (rather than reassigned) as an explicit strategy employed by the recovering school principals to improve outcomes. Furthermore, the application of short-run academic improvement strategies also creates certain doubts as to the quality of learning and the desirability of these practices in Chilean schools. Other studies are needed in order to explore the qualitative effects of the policies and practices developed by schools that are under pressure from SEP.

Another relevant finding from this study is that the main changes at recovering schools were made on the level of teaching policies, but not practices. In fact, only two differences were detected among the 25 practices analyzed: greater use of multiple-choice tests and project reports as tools for student evaluation. There appears to be no clear relationship between these two practices. In general, the pedagogical activities carried out by teachers at different schools seem to be, on the contrary, rather homogenous.

This, along with the fact that two out of every three teachers surveyed did not know their school's ranking, suggests that most of these changes are planned and applied only at the administrative level, without involving teachers in the process. Otherwise, teachers would know their schools' classification and would be aware of the sanctions that the school would face if it did not show significant improvement. This reveals a problem in the implementation of the SEP Law in schools, which could explain the absence of significant differences in teaching practices among recovering and emerging schools. Moreover, making top down changes in the school top-down could be problematic in light of international evidence. Recent research suggests that the success stories of school accountability, in terms of improvement, occur in schools where the entire educational community, especially teachers, participates in the process in order to improve its performance (e.g. Mintrop, 2012).

Acknowledgments

We thank the comments from Santiago Cueto, Roberto Pinto and Pablo González, and the participants in the seminar to disseminate the project results, organized by PREAL in Lima, Perú.

Funding

This research was funded by the Fund for Research in Education of PREAL (Programa de Promoción de la Reforma Educativa en América Latina y el Caribe).

Notes

1. Details on the methodology used to classify students as priority can be found in [] and [] (2013).
2. [] and [] (2013) find that 84% of municipal and private voucher schools have decided to participate in SEP. However, there are important differences in the participation rate of these schools. While almost all municipal schools (99%) participate, only 61% of private voucher schools have decided to enter the system.
3. For example, the SEP Law requires that all participating schools must ban selection practices by socioeconomic or academic criteria between pre-K and sixth grade. Schools can only use academic assessments to select students in seventh to twelfth grade. In addition, schools are not allowed to charge fees to families of priority students and are required to make public how they spend the additional SEP resources.
4. The final classification of each school was published in September 2011, by letter addressed to the school principal. The letter provided details on the school´s ranking and how the school can appeal the classification and its obligations according to the law.

5. There is evidence that shows high volatility in standardized test scores from year to year in Chile, which makes producing a meaningful ranking of schools a challenge (e.g. Mizala, Romaguera, & Urquiola, 2007). If schools leave the recovering category due to statistical noise rather than substantial changes in their practices, it could be a disincentive to implement changes. To address this potential problem, the design of SEP contemplates using the results of fourth-grade standardized tests the last three years prior to the classification.
6. Several studies in Chile have evaluated the impact of school-based accountability programs through Regresssion Discontinuity Designs. For example, Chay, McEwan, and Urquiola (2005) evaluate Chile's 900 Schools Program (P-900), which allocated resources based on cutoffs in schools' mean test scores. Also see Mizala and Urquiola (2013) for an evaluation of the SNED program, which seeks to identify effective schools selecting them from "homogeneous groups" of comparable institutions.
7. To view the details of the other estimation methods in the context of n assignment variables see also Wong, Steiner, and Cook (2013).
8. $Z_i < 0$ implies $R_i < r_c$ and $M_i < m_c$, and then unit i is assigned to the treatment condition.
9. SIMCE is the oldest evaluation system in Latin America. It has functioned annually since 1988 (although its origin goes back to the early 1980s). The SIMCE tests are applied to all students in the second, fourth, sixth, eighth, and tenth grades at the national-level. SIMCE also gathers detailed information about teachers, students, and parents.
10. This level was chosen because the current school classification methodology in the SEP law only considers fourth-grade SIMCE test results. Therefore, the impact of the threat on teaching policies and practices should be more significant at this level.
11. The formula incorporates as one of its parameters the Minimum Detectable Effect (MDE) that we attempt to estimate. For our calculation, this effect was set at 0.59 standard deviations. Although it is a large size, there are limitations in the program design (the sample considers the recovering schools universe) and budget constraints that impede the consideration of a larger sample.
12. In order to test for potential biases in the sample, a Logit model was estimated to explain the probability of rejection. In general, the results show that there are no significant differences between schools that decided to participate and those that didn't. These results are available upon request.
13. The MDE associated with this sample size is 0.72 standard deviations.
14. The complete table is available upon request.
15. For the implementationof the test, all schools in Greater Santiago participating in the SEP Law are considered (689 schools).
16. The results of this estimations are available upon request.
17. The results, which are not presented in this paper, are available upon request.
18. Results from models with polynomials of degree 3 and higher were discarded because, in most cases, the magnitude of the estimate is too large, probably due to overfitting. On the other hand, a high degree of colinearity was detected in these higher order models, thereby reducing the accuracy of the estimates. Estimation results are available upon request.
19. The Inicia test is an assessment of graduates from undergraduate teacher education programs, which aims to provide information on the quality of teacher training. Despite being voluntary, in 2010 it was taken by graduates in 43 of the 59 schools of education in Chile. In total, 2,111 graduates were evaluated. The assessment includes four tests: i) general knowledge of elementary education; ii) pedagogical and theoretical knowledge; iii) written communication skills; and iv) skills for the management of information technology and communication. The variable used in this paper is the percentage of correct answers obtained by students from their undergraduate school of education on the general knowledge test.

References

Ahn, T., & Vigdor, J. (2009).*Does No Child Left Behind have teeth? Examining the impact of federal accountability sanctions in North Carolina*. Unpublished manuscript.

Booher-Jennings, J. (2005). Below the Bubble: Educational Triage and the Texas Accountability System. *American Educational Research Journal, 42*, 231–268. doi: 10.3102/00028312042002231

Brady, R. (2003). *Can Failing Schools be fixed?* Washington, DC: Thomas B. Fordham Foundation.

Carnoy, M., & Loeb, S. (2002). Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis*, *24,* 305–331. doi: 10.3102/01623737024004305

Chakrabarti, R. (2008). *Impact of Voucher Design on Public School Performance: Evidence from Florida and Milwaukee Voucher Program*. (Federal Reserve Bank of New York Staff ReportNo. 315). Retrieved from http://www.newyorkfed.org/research/staff_reports/sr315.pdf

Chay, K., McEwan, P., & Urquiola, M. (2005). The Central Role of Noise in Evaluating Interventions that Use Test Scores to Rank Schools. *The American Economic Review, 95*, 1237-1258. doi**:** 10.1257/0002828054825529

Cullen, J. B., & Reback, R. (2006). Tinkering toward accolades: School gaming under a performance accountability system. *NBER Working Paper* No. 12286. Retrieved from http://www.nber.org/papers/w12286

Deming, D., Cohodes, S., Jennings, J., & Jencks, C. (2013). School accountability, postsecondary attainment and earnings. *NBER Working Paper* No 19444. Retrieved from http://www.nber.org/papers/w19444

Elliot, R., &Borko, H. (1999). Hands-On Pedagogy vs. Hands-Off Accountability: Tensions Between Competing Commitments for Exemplary Math Teachers in Kentucky. *Phi Delta Kappan*, *80*, 394–400. Retrieved from http://www.jstor.org/stable/20439455

Elmore, R. F., Abelman, C. H., & Fuhrman, S. H. (1996). The new accountability in state education reform: From process to performance. In H. F. Ladd (Ed.), *Holding schools accountable: Performance-based reform in education* (pp. 65-98). Washington, DC: Brookings Institution.

Figlio, D. N. (2006). Testing, crime and punishment. *Journal of Public Economics*, *90*, 837-851.doi: 10.1016/j.jpubeco.2005.01.003

Figlio, D. N., & Getzler, L. S. (2006). Accountability, Ability and Disability: Gaming the System? In T.J. Gronberg, & D.W. Jansen (Eds.), *Improving School Accountability (Advances in Applied Microeconomics, Volume 14)* (pp.35-49). Bingley, UK: Emerald Group Publishing Limited

Figlio, D.N., & Loeb, S. (2011). School Accountability. In E. A.Hanushek, S. Machin, & L. Woessmann (Eds.), *Economics of Education* (pp. 383-421). The Netherlands: North Holland.

Figlio, D. N., & Lucas, M. E. (2004). Whats in a Grade? School Report Cards and the Housing Market. *The American Economic Review*, *94*, 591-604. doi: 10.1257/0002828041464489

Figlio, D. N., & Rouse, C. E. (2006). Do Accountability and Voucher Threats Improve Low-Performing Schools? *Journal of Public Economics, 90,* 239-255. doi: 10.1016/j.jpubeco.2005.08.005

Gill, B., Lockwood, J., Martorell, F., Setodji, C. M., Booker, K., Vernez, G., Birman, B., & Garet, M. S. (2009). *An Exploratory Analysis of Adequate Yearly Progress, Identification for Improvement, and Student Achievement in Two States and Three Cities*.(Technical Report U.S. Department of Education). Retrieved from http://www2.ed.gov/rschstat/eval/disadv/rd-ayp/report.pdf

Goldhaber, D., &Hannaway, J. (2004). Accountability with a kicker: Observations on the Florida A+ accountability plan. *Phi Delta Kappan, 85*, 598-605. Retrieved from http://www.jstor.org/stable/20441651

Greene, J. P., & Winters, M. A. (2003). When Schools Compete: The Effect of Vouchers on Florida Public School Achievement. *Manhattan Institute Education Working Paper* No. 2. Retrieved from http://www.manhattan-institute.org/pdf/ewp_02.pdf

Hamilton, L., Stecher, B., Marsh, J., McCombs, J., Robyn, A., Russel, J. N., & Barney, H. (2007). *Implementing standards-based accountability under No Child Left Behind: Responses of superintendents, principals, and teachers in three states*. Santa Monica, CA: RAND. Retrieved from http://www.rand.org/content/dam/rand/pubs/monographs/2007/RAND_MG589.pdf

Hannaway, J., & Cohodes, S. (2007). Trouble Even in Choice Paradise: NCLB in Miami-Dade County Public Schools. In F. M. Hess, & C. E. Finn (Eds.), *No remedy left behind: Lessons from a half-decade of NCLB* (pp. 244-264). Washington DC: American Enterprise Institute Press.

Hannaway, J., & Hamilton, L. (2007). *Performance-Based Accountability Policies: Implications for School and Classroom Practices*. Washington, DC: The Urban Institute and RAND Corporation. Retrieved from http://www.urban.org/UploadedPDF/411779_accountability_policies.pdf

Hanushek, E., & Raymond, M. (2003). Lessons about the Design of State Accountability Systems. In E. Peterson, & M. West (Eds.), *No Child Left Behind? The Politics and Practice of Accountability* (pp. 126-151). Washington, DC: Brookings.

Hanushek, E. A., & Raymond, M. E. (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management, 24*, 297-327. doi: 10.1002/pam.20091

Hirschman, A. O. (1970). *Exit, voice, and loyalty: Responses to decline in firms, organizations, and states*. Cambridge, MA: Harvard University Press.

Imbens, G. W., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics, 142*, 615–635. doi: 10.1016/j.jeconom.2007.05.001

Imbens, G., & Zajonc, T. (2011). Regression Discontinuity Design with Multiple Forcing Variables. Unpublished manuscript.

Jacob, B. A. (2005). Accountability, incentives and behavior: the impact of high-stakes testing in Chicago Public Schools. *Journal of Public Economics, 89*, 761-796. doi: 10.1016/j.jpubeco.2004.08.004

Jacob, B. A., & Levitt, S. D. (2003). Rotten apples: An Investigation of the Prevalence and Predictors of Teacher Cheating. *The Quarterly Journal of Economics, 118*, 843-877. doi: 10.1162/00335530360698441

Koretz, D., Stecher, B., Klein, S., & McCaffrey, D. (1994). The Vermont Portafolio Assessment Program: Findings and Implications. *Educational Measurament Issues and Practices 13(3)*, 5-16.doi: 10.1111/j.1745-3992.1994.tb00443.x

Koschoreck, J. W. (2001). Accountability and Educational Equity in the Transformation of an Urban District. *Education and Urban Society, 33*, 284-304. doi: 10.1177/0013124501333004

Ladd, H.F, & Glennie, E. (2001). A replication of Jay Greene's voucher effect study using North Carolina data. In M. Carnoy (Ed.), Do School Vouchers Improve Student Performance? (pp. 49-52). Washington, DC: Economic Policy Institute.

Ladd, H. F., & Zelli, A. (2002). School-based accountability in North Carolina: The responses of school principals. *Educational Administration Quarterly, 38*, 494-529. doi: 10.1177/001316102237670

Lee, D. S. (2008). Randomized experiments from non-random selection in U.S. House elections. *Journal of Econometrics, 142*, 675-697. doi: 10.1016/j.jeconom.2007.05.004

Lee, D. S., & Lemieux, T. (2010). Regression Discontinuity Designs in Economics. *Journal of Economic Literature, 48*, 281–355. doi: 10.1257/jel.48.2.281

Martorell, F. (2005). *Do high school graduation exams matter? Evaluating the effects of exit exam performance on student outcomes*. Unpublished manuscript.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics, 142*, 698-714. doi: 10.1016/j.jeconom.2007.05.005

Mintrop, H. (2012). Bridging accountability obligations, professional values and (perceived) student needs with integrity. *Journal of Educational Administration, 50*, 695-726.doi: 10.1108/09578231211249871

Mizala, A. & Urquiola, M. (2013). School markets: The impact of information approximating schools' effectiveness. *Journal of Development Economics 103*, 313–335. doi: 10.1016/j.jdeveco.2013.03.003

Mizala, A., Romaguera, P., & Urquiola, M. (2007). Socioeconomic status or noise? Tradeoffs in the generation of school quality information. *Journal of Development Economics*, *84*, 61-75. doi: 10.1016/j.jdeveco.2006.09.003

Neal, D., & Schanzenbach, D. W. (2007). Left behind by design: Proficiency counts and test-based accountability. *NBER Working Paper* No. 13293. Retrieved from http://www.nber.org/papers/w13293

O'Day, J. (2002). Complexity, Accountability, and School Improvement. *Harvard Educational Review*, 72, 293-329. Retrieved from http://her.hepg.org/content/021q742t8182h238/

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. doi: 10.1016/j.jeconom.2010.12.008

Pedulla, J., Abrams, L., Madaus, G., Russell, M., Ramos, M., & Miao, J. (2003). *Perceived Effects of State-Mandated Testing Programs on Teaching and Learning: Findings from a National Survey of Teachers*. Boston, MA: National Board on Education Testing and Public Policy.

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

Reardon, S.F, & Robinson, J. (2012). Regression Discontinuity Designs With Multiple Rating-Score Variables. *Journal of Research on Educational Effectiveness, 5*, 83-104. doi: 10.1080/19345747.2011.609583

Reardon, S. F., Arshan, N., Atteberry, A., & Kurlaender, M. (2010). Effects of Failing a High School Exit Exam on Course Taking, Achievement, Persistence, and Graduation. *Educational Evaluation and Policy Analysis, 32*, 498–520. doi: 10.3102/0162373710382655

Reback, R., Rockoff, J., & Schwartz, H. L. (2011). Under pressure: Job security, resource allocation, and productivity in schools under NCLB. *NBER Working Paper* No. 16745. Retrieved from http://www.nber.org/papers/w16745

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, 33*, 267–292. doi: 10.3102/0162373711407912

Rockoff, J., & Turner, L. (2010). Short-Run Impacts of Accountability on School Quality.*American Economic Journal: Economic Policy*, 2: 119-47.doi: 10.1257/pol.2.4.119

Romberg, T., Zarinnia, A., & Williams, S. (1989). The Influence of Mandated Testing on Mathematics Instruction: Grade 8 Teachers' Perceptions. Madison, WI: National Center for Research in Mathematical Sciences Education.

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.doi: 10.1257/pol.5.2.251

Scheurich, J. J., Skrla, L., & Johnson, J. F. (2000). Thinking Carefully About Equity and Accountability. *Phi Delta Kappan, 82*, 293-299. Retrieved from http://www.jstor.org/stable/20439884

Schochet, P. Z. (2009). Statistical Power for Regression Discontinuity Designs in Education Evaluations. *Journal of Educational and Behavioral Statistics, 34*, 238–266. doi: 10.3102/1076998609332748

Shepard, L., & Dougherty, K. (1991, April). Effe*cts of High-stakes Testing on Instruction*.Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

Smarick, A. (2010). The turnaround fallacy. *Education Next, 10*, 20-26. Retrieved from http://educationnext.org/the-turnaround-fallacy/

Smith, M. L., & Rottenberg, C. (1991). Unintended consequences of external testing in elementary schools. *Educational Measurement: Issues and Practice, 10*(4), 7-11. doi: 10.1111/j.1745-3992.1991.tb00210.x

Spreng, C. (2005). Policy Options for Interventions in Failing Schools. Santa Monica, CA: Rand Corporation.

Stecher, B. M., & Barron, S. (2001). Unintended consequences of test-based accountability when testing in "milepost" grades. *Educational Assessment, 7*, 259-281. doi: 10.1207/S15326977EA0704_02

Stecher, B. M., Barron, S. L., Chun, T., & Ross, K. (2000). The effects of the Washington state education reform on schools and classrooms. Santa Monica, CA: RAND Corporation.

West, M. R., & Peterson, P. E. (2006). The efficacy of choice threats within school accountability systems: Results from legislatively induced experiments. *The Economic Journal, 116*, C46-C62. doi: 10.1111/j.1468-0297.2006.01075.x

Wolf, S., Borko, H., McIver, M., & Elliott, R. (1999). No Excuses: School Reform Efforts in Exemplary Schools of Kentucky. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

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. doi: 10.3102/1076998611432172

Appendix. Construction of the binding-score (Z)

Table A1 shows the eight variables that define if a school is classified as recovering or emerging. In order to have all the variables on the same scale, each variable was centered on the respective cutoff and then divided by standard deviation. For example, for $psimce_{2010}$ we construct a new variable,$psimce^{zc}_{2010}$ defined as:

$$psimce^{zc}_{2010} = \frac{psimce_{2010}-220}{\sigma_{psimce_{2010}}}.$$

First, we calculate the maximum between$psimce$and$p250$for each year:

$$Z_1 = max(psimce^{zc}_{2010}, p250^{zc}_{2010})$$

$$Z_2 = max(psimce^{zc}_{2009}, p250^{zc}_{2009})$$

$$Z_3 = max(psimce^{zc}_{2008}, p250^{zc}_{2008}).$$

Then, we calculate the second maximum between Z_1, Z_2 andZ_3:

$$Z_4 = secondmax(Z_1, Z_2, Z_3).$$

Thus, Z_4 indicates if a school is classified as recovering according to SIMCE results. If$Z_4 < 0$, then $psimce_{i,t} < 220$ and $p250_{i,t} < 20\%$ in two years, and therefore the school meets the requirements to be classified as recovering. The opposite is true when $Z_4 \geq 0$. To incorporate the Education Quality Index, we calculate the minimum between Z_4 and ICE^{zc}:

$$Z_5 = min(Z_4, ICE^{zc}).$$

Finally, a set of schools that meet (so far) the conditions to be classified as recovering ($Z_5 < 0$) are considered emerging because fewer than 20students took the SIMCE test ($nsimce < 20$). To incorporate this condition, we calculate the maximum between Z_5 and minus $nsimce^{zc}$.

$$Z = max(Z_5, -nsimce^{zc})$$

This variable (binding-score) perfectly determines treatment assignment. If $Z_i < 0$, a school is classified asrecovering, while if$Z_i \geq 0$, it is classified as emerging. Table A2 shows which of the eight assignment variables correspond to Z for each school. The results indicate thatmostschools are onthe frontierdefined byICE^{zc}_i (72.3%). Anotherimportant grouparethose with $Z_i = nsimce^{zc}_i$ (11.2%). The restis locatedon the frontierdefined by SIMCEtest results (16.5%).

[Table A1]

[Table A2]

Figures

Figure 1: Estimated density (Z_i)

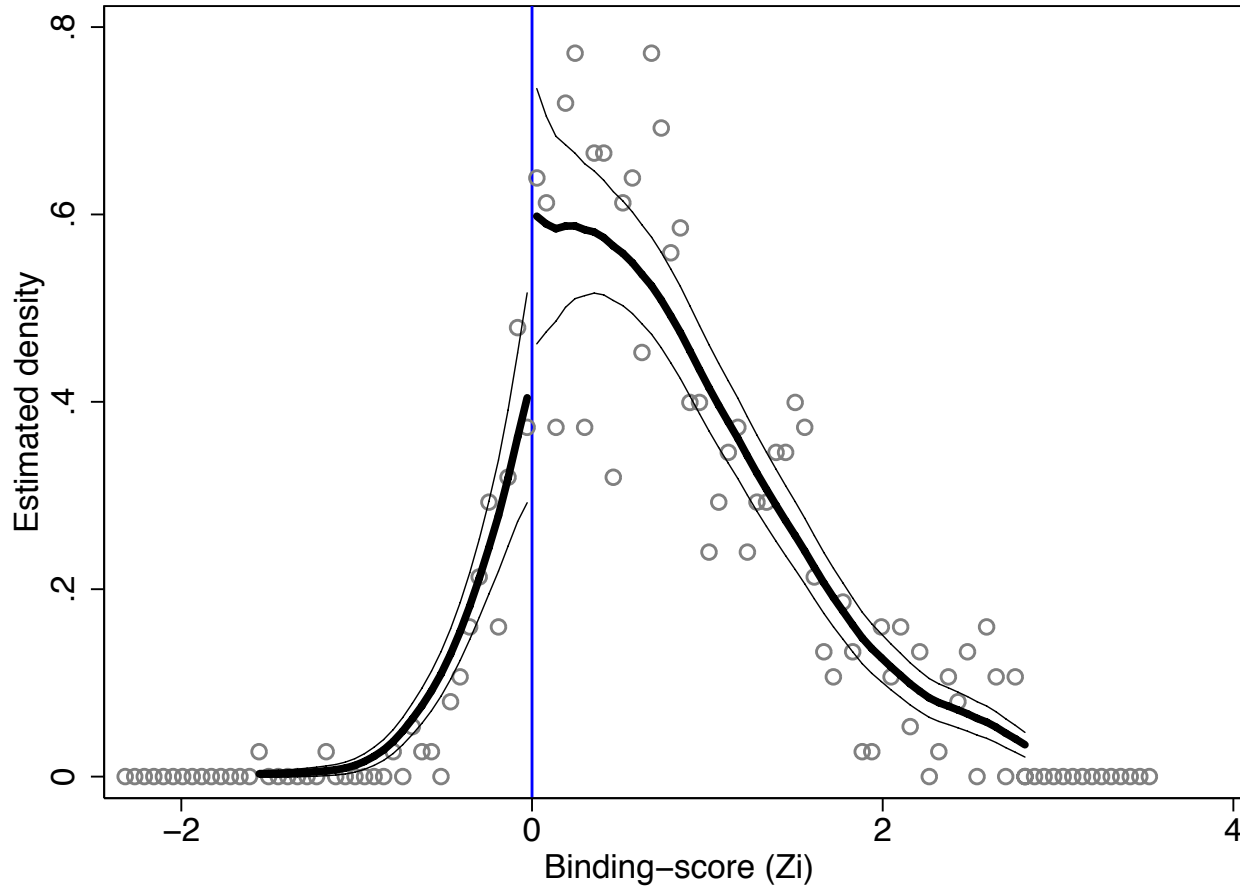


Figure 2: Graphical results for four key findings

Figure 2 (a): Use of personal tutors after class for low-achievement students

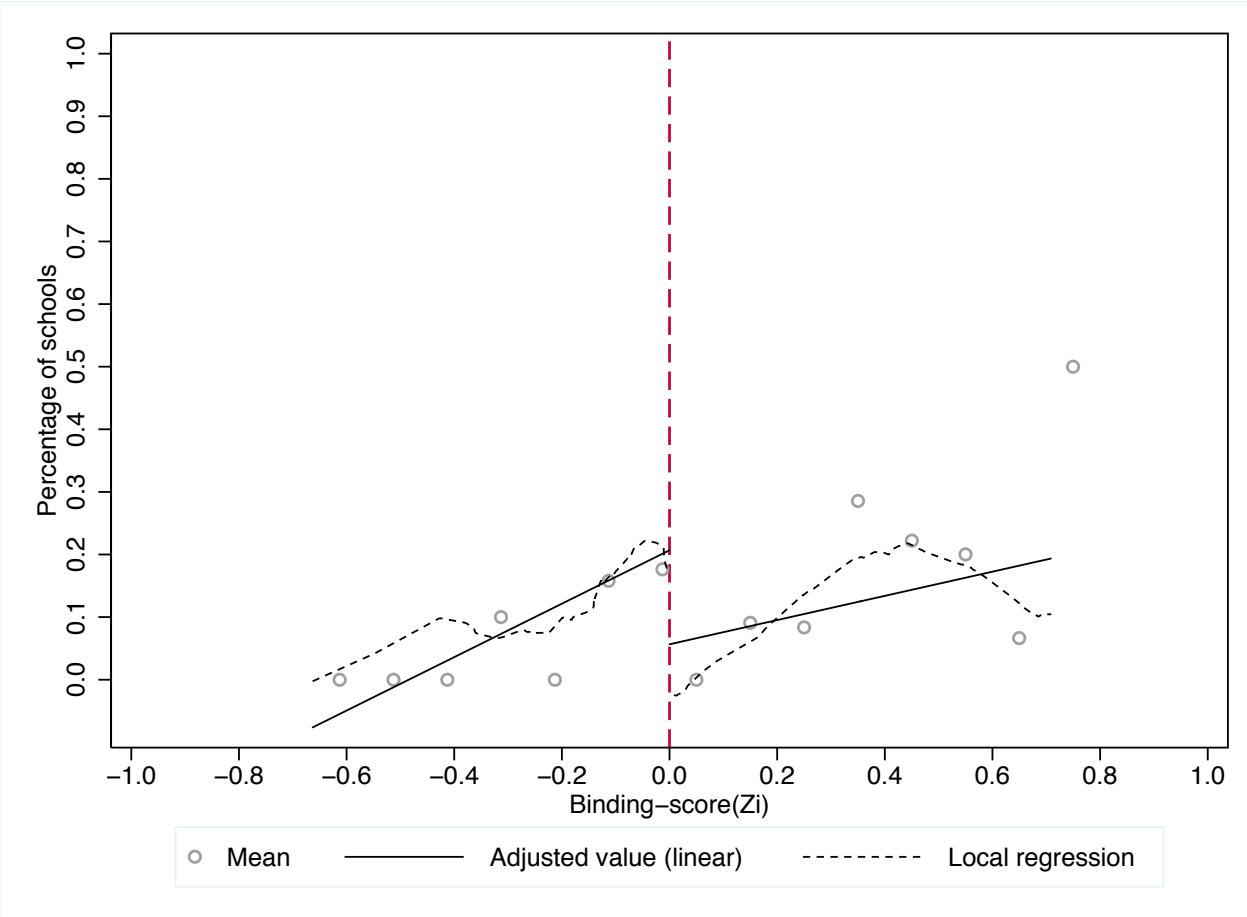


Figure 2 (b): Mandatory attendance at training as a sanction for low-performing teachers

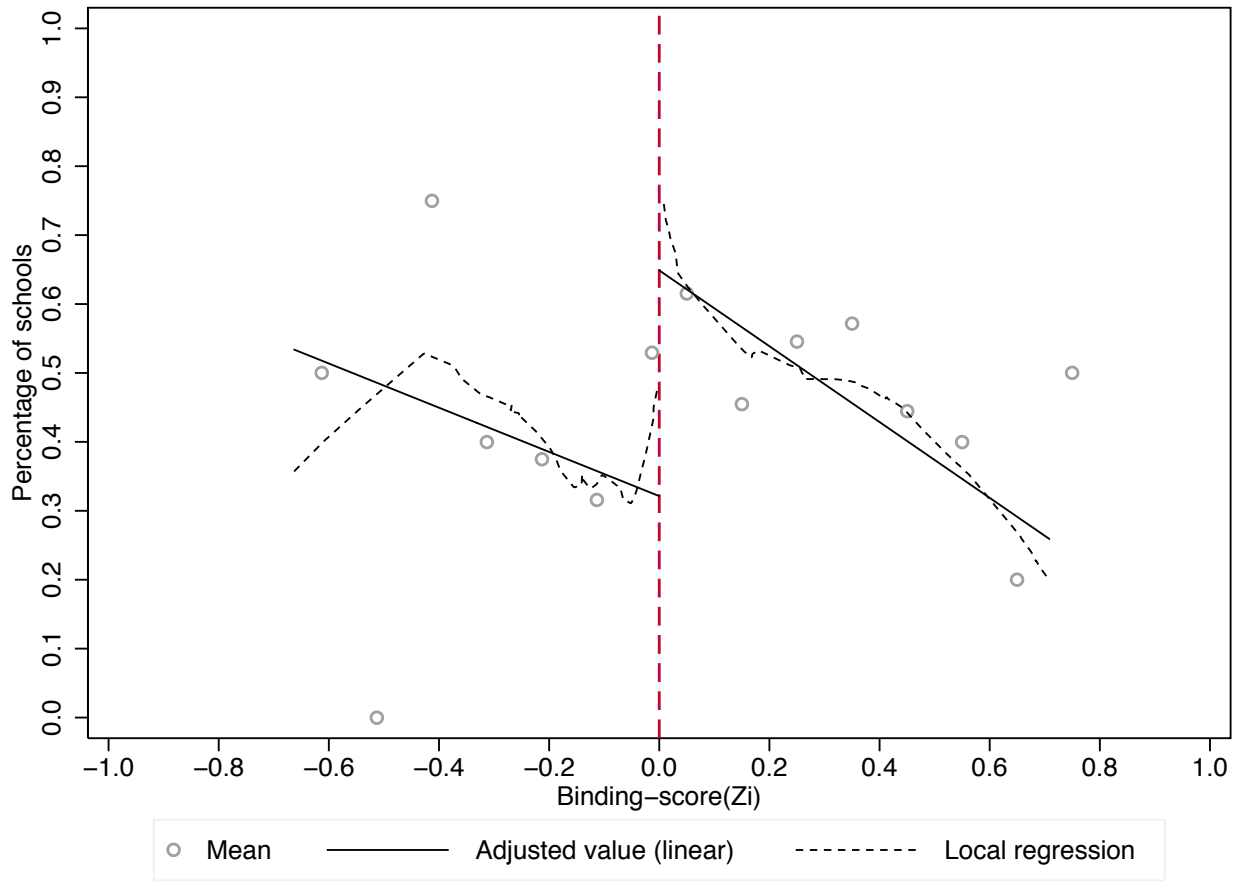


Figure 2 (c): The teacher observed one of his/her colleagues conducting a class

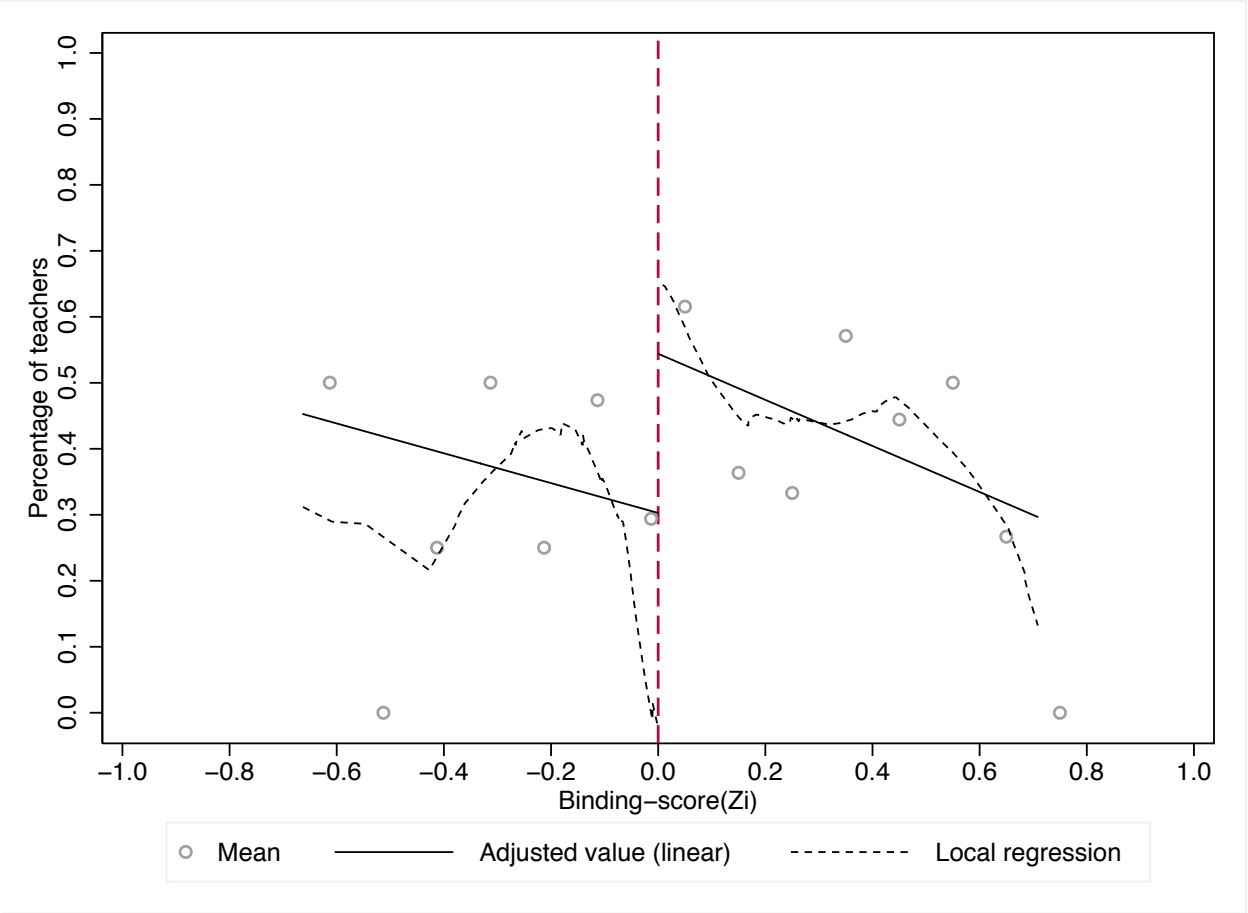
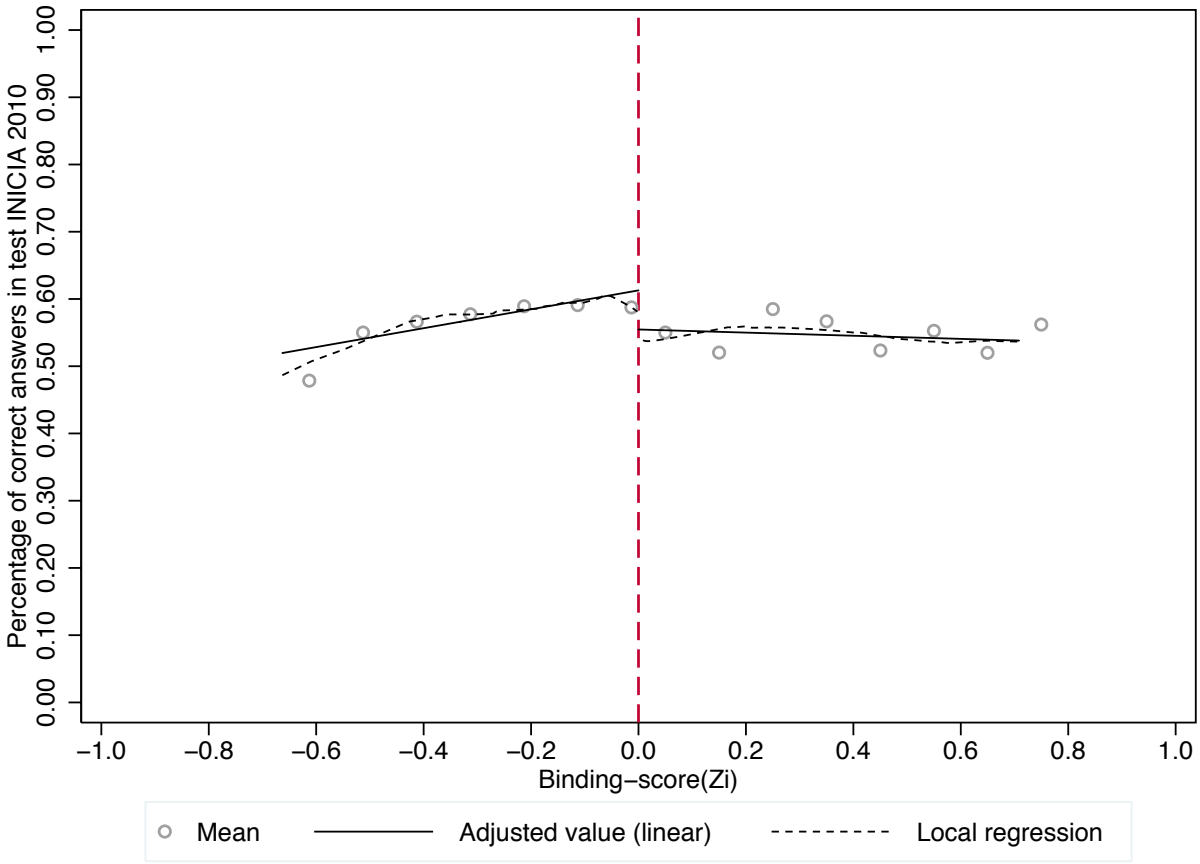


Figure 2 (d): Inicia test results (2010) from graduated teacher institution



Tables

Table 1. Summary of the sample's descriptive statistics

	Recovering		Emerging		Difference in means/proportions test	
Domain/Variable	Mean	N	Mean	N	t/z	p-value
Internal teacher evaluations						
There are minimum performance goals in SIMCE test at school (1=Yes)	73.6%	53	78.5%	79	0.651	0.515
There is an internal teacher assessment at school (1=Yes)	38.5%	52	55.1%	78	1.863	0.062*
Policies focused on low-performing teachers						
Provision of a support plan for classes (1=Yes)	40.7%	54	57.0%	79	1.837	0.066*
SIMCE preparation						
In the last four weeks the teacher:						
Taught using similar exercises to SIMCE (1=every day or almost every day)	63.0%	54	63.8%	80	0.093	0.926
Taught how to answer multiple-choice tests (1=every day or almost every day)	55.6%	54	67.5%	80	1.402	0.161
Internal evaluations applied to students						
In the last four weeks the teacher:						
Evaluated students through a multiple choice test (1=yes)	90.7%	54	97.5%	80	1.725	0.085*
Evaluated students through a SIMCE test (1=yes)	72.2%	54	73.8%	80	0.196	0.845
Parent Involvement						
In the last four weeks:						
Some guardian requested a face to face meeting with the teacher (1=Yes)	24.5%	53	43.0%	79	2.177	0.029**
Qualitative SEP evaluation						
Do you know the Law SEP? (1=Yes)	92.6%	54	95.0%	80	0.577	0.564
Do you know the SEP classification? (1=Yes)	98.0%	51	97.1%	70	-0.313	0.754
Distribution of weekly class hours among different subjects						
Percentage of weekly class hours devoted to areas covered by SIMCE	58.8%	41	59.6%	61	0.593	0.554
School characteristics						
Average years of schooling of the mother (2010)	9.0	54	9.7	80	4.040	0.000***
Average household income (2010 USD)	386	54	426	80	2.456	0.015**
Percentage of students who have more than one hundred books at home (2010)	2.3%	54	4.6%	80	2.638	0.009***
Percentage of students who have repeated a grade (2010)	31.3%	54	23.1%	80	-3.289	0.001***
Percentage of parents who attend meetings (2010)	78.2%	54	80.8%	80	1.563	0.120
Total per student expenditure (2009 USD)	94	54	91	80	-1.038	0.301
Student/teacher ratio (2010)	19.9	54	19.5	80	-0.405	0.687

*** p < 0.01; ** p < 0.05; * p < 0.1

Source: Authors' calculations based on teacher survey and SIMCE databases.

Table 2. McCrary test

Variable	Test t
$psimce_{2010}$	-1.45
$psimce_{2009}$	0.45
$psimce_{2008}$	1.02
$p250_{2010}$	1.76*
$p250_{2009}$	0.67
$p250_{2008}$	2.25**
ICE	0.78
nsimce	1.60
Binding-score (Z_i)	1.55

Notes : To simplify the notation, subscripts are omitted, but each variable is centered on its respective cutoff and divided by its standard deviation.

*** p < 0.01; ** p < 0.05; * p < 0.1.

Source : Authors' calculation

Table 3. Average treatment effect on different outcomes (Bandwidth = 0.71)

	N	Polynomial order 1		Polynomial order 2	
		No covariates	With covariates[a]	No covariates	With covariates[a]
School Policies					
Attendance at training as a reward for high-performing teachers (1=Yes)	120	-0.299*	-0.350**	-0.193	-0.316
		(0.158)	(0.173)	(0.223)	(0.236)
Mandatory attendance at training as a sanction for low-perfoming teachers (1=Yes)	128	-0.400***	-0.367**	-0.407%*	-0.364
		(0.149)	(0.160)	(0.211)	(0.221)
Personal tutor after school for low-achieving students (1=Yes)	128	0.206	0.264*	0.203	0.243
		(0.130)	(0.149)	(0.183)	(0.205)
External specialist observed classes (1=every day or almost every day in the last four weeks)	128	-0.212	-0.195	-0.468**	-0.466**
		(0.146)	(0.158)	(0.204)	(0.214)
The teacher watched one of his colleagues teach a class (1=every day or almost every day in the last four weeks)	128	-0.250	-0.229	-0.318	-0.263
		(0.151)	(0.164)	(0.211)	(0.224)
There is an internal teacher assessment at school (1=Yes)	124	-0.176	-0.259	-0.475**	-0.558**
		(0.157)	(0.166)	(0.217)	(0.225)
Teacher Practices					
Teacher evaluated students using a research project in the last four weeks (1=Yes)	128	0.073	0.133	0.274*	0.322*
		(0.117)	(0.131)	(0.163)	(0.178)
Teacher evaluated students using a multiple choice test in the last four weeks (1=at least 2 times per week)	128	0.192	0.223	0.332	0.368*
		(0.130)	(0.149)	(0.184)	(0.205)
Teacher Characteristics					
Contract hours in classroom (chronological hours)	123	1.229	2.248	3.050	4.350
		(2.064)	(2.244)	(2.874)	(3.049)
Teacher has a graduate degree (1=Yes)	128	0.066	0.076	0.361	0.354
		(0.153)	(0.173)	(0.211)	(0.235)
Teacher does not have an undergraduate area of specialization (1=Yes)	128	-0.210	-0.268	-0.209	-0.226
		(0.154)	(0.172)	(0.218)	(0.238)
Undergraduate institution exit competency exam results (% correct answers)	69	0.064*	0.067*	0.064	0.121**
		(0.036)	(0.039)	(0.050)	(0.052)
Years of teaching experience (number of years)	127	6.038	4.845	10.866*	10.360*
		(3.999)	(4.070)	(5.599)	(5.519)

Notes: Standard deviation in parentheses.

*** p < 0.01; ** p < 0.05; * p < 0.1.

[a] The regression with covariates included characteristics of school students, variables measuring the availability of resources in the school, faculty characteristics and a set of variables that are used as a proxy for management type of establishment management team.

Source : Author's calculations based on survey of teachers.

Table 4. Pretreatment variables included in the regressions

Variable	Year	Source
Students Characteristics		
Average mother's years of schooling	2010	Parent Questionnaire SIMCE 4th grade
Average household income	2010	Parent Questionnaire SIMCE 4th grade
Percentage of students with more than 100 books at home	2010	Parent Questionnaire SIMCE 4th grade
Percentage of students who have failed a grade	2010	Parent Questionnaire SIMCE 4th grade
School Resources		
Total spending per student[a]	2009	Average Student Attendance and Public Voucher Databases
Student/teacher ratio	2010	Student Enrollment and Teacher Census Databases
School Type (1=Public School)	2011	School Directory
School has an extended school day program (JEC)[b]	2009	Average Student Attendance and Public Voucher Databases
Faculty Characteristics		
Average age	2010	Teacher Census Database
Average hours of contract	2010	Teacher Census Database
Average work experience	2010	Teacher Census Database
Percentage of classroom teachers	2010	Teacher Census Database
Percentage of teachers from Unidad Técnico Pedagógica (UTP)[c]	2010	Teacher Census Database
Percentage of teachers without a degree in education	2010	Teacher Census Database
Percentage of female teachers	2010	Teacher Census Database
Managment Team		
Management Team assesses the teacher's impact (1=Yes)	2010	Teacher Questionnaire SIMCE 4th grade
Management team supports teachers to improve (1=Yes)	2010	Teacher Questionnaire SIMCE 4th grade
Management Team defines clear learning goals (1=Yes)	2010	Teacher Questionnaire SIMCE 4th grade
Management team promotes teacher development (1=Yes)	2010	Teacher Questionnaire SIMCE 4th grade

[a] Total spending per student includes the resources that the school receives from state subsidies and private contributions of families (monthly fees).

[b] JEC is a program funded through subsidies and direct contributions from the State that seeks to increase the time students spend in school by approximately 30% annually. This program is targeted for subsidized schools (both public and private).

[c] The Unidad Técnico Pedagógica (UTP) is the school agency responsible for coordinating, advising, and evaluating technical teachers pedagogical functions such as Educational and Vocational Guidance, Curriculum Planning, Educational Supervision, and Evaluation of Learning. It aims to optimize the development of pedagogical and technical processes, and become the main support leading changes and improvement actions within the school.

Table 5. Average treatment effect on different outcomes (Bandwidth = 0.32)

	N	Polynomial order 1		Polynomial order 2	
		No covariates	With covariates[a]	No covariates	With covariates[a]
School Policies					
Attendance at training as a reward for high-performing teachers (1=Yes)	74	-0.192	-0.208	-0.200	-0.176
		(0.213)	(0.244)	(0.330)	(0.372)
Mandatory attendance at training as a sanction for low-perfoming teachers (1=Yes)	79	-0.435**	-0.258	-0.345	-0.239
		(0.202)	(0.214)	(0.304)	(0.308)
Personal tutor after school for low-achieving students (1=Yes)	79	0.162	0.338*	0.218	0.164
		(0.169)	(0.198)	(0.254)	(0.284)
External specialist observed classes (1=every day or almost every day in the last four weeks)	79	-0.634***	-0.517**	-0.354	-0.295
		(0.179)	(0.200)	(0.255)	(0.278)
The teacher watched one of his colleagues teach a class (1=every day or almost every day in the last four weeks)	79	-0.395*	-0.450**	-0.688**	-0.938***
		(0.202)	(0.224)	(0.301)	(0.307)
There is an internal teacher assessment at school (1=Yes)	76	-0.320	-0.475**	-0.367	-0.600**
		(0.210)	(0.202)	(0.313)	(0.293)
Teacher Practices					
Teacher evaluated students using a research project in the last four weeks (1=Yes)	79	0.163	0.287	0.385*	0.460*
		(0.145)	(0.176)	(0.215)	(0.252)
Teacher evaluated students using a multiple choice test in the last four weeks (1=at least 2 times per week)	79	0.292	0.299	0.305	0.108
		(0.159)	(0.183)	(0.239)	(0.261)
Teacher Characteristics					
Contract hours in classroom (chronological hours)	74	4.108	3.777	7.148*	11.142**
		(2.779)	(3.171)	(4.154)	(4.286)
Teacher has a graduate degree (1=Yes)	79	0.273	0.266	0.674**	0.732**
		(0.207)	(0.237)	(0.299)	(0.338)
Teacher does not have an undergraduate area of specialization (1=Yes)	79	-0.491**	-0.544**	-0.590*	-0.488
		(0.202)	(0.225)	(0.305)	(0.325)
Undergraduate institution exit competency exam results (% correct answers)	38	0.067	0.155**	0.088	0.148
		(0.052)	(0.057)	(0.084)	(0.117)
Years of teaching experience (number of years)	78	8.886	10.234*	8.421	6.934
		(5.405)	(5.717)	(8.148)	(8.201)

Notes: Standard deviation in parentheses.

*** p < 0.01; ** p < 0.05; * p < 0.1.

[a] The regression with covariates included characteristics of school students, variables measuring the availability of resources in the school, faculty characteristics and a set of variables that are used as a proxy for management type of establishment management team.

Source : Author's calculations based on survey of teachers.

Table A1. Variables that define if a school is classified as recovering or emerging

Variable	Description	Cutoff
psimce $_{2010}$	School average SIMCE Fourth Grade 2010 score	220
psimce $_{2009}$	School average SIMCE Fourth Grade 2009 score	220
psimce $_{2008}$	School average SIMCE Fourth Grade 2008 score	220
p250 $_{2010}$	School average proportion of students who have scored over 250 points in SIMCE Fourth Grade 2010	20%
p250 $_{2009}$	School average proportion of students who have scored over 250 points in SIMCE Fourth Grade 2009	20%
p250 $_{2008}$	School average proportion of students who have scored over 250 points in SIMCE Fourth Grade 2008	20%
Education Quality Index (ICE)	Index that combines average SIMCE score of three years[a] (70%) with complementary indicators (30%)	10th percentile
nsimce[b]	Average number of students who take SIMCE	20

Notes : The first 6 variables correspond to the average of the averages in each area assessed in SIMCE (Language, Math and Science).
a. Average SIMCE score used in the construction of ICE corresponds to the average of psimce $_{2010}$, psimce $_{2009}$ and psimce $_{2008}$.
b. nsimce is obtained by averaging the number of students who took SIMCE in each year and area assessed.
Source: Ministry of Education.

Table A2. Binding-score (Z)

Binding-score	Recovering		Emerging		Total	
	N	%	N	%	N	%
psimce $_{2010}$	1	1.1%	0	0.0%	1	0.1%
psimce $_{2009}$	0	0.0%	2	0.3%	2	0.3%
psimce $_{2008}$	0	0.0%	3	0.5%	3	0.4%
p250 $_{2010}$	3	3.4%	30	5.0%	33	4.8%
p250 $_{2009}$	5	5.7%	39	6.5%	44	6.4%
p250 $_{2008}$	1	1.1%	30	5.0%	31	4.5%
ICE	46	52.9%	452	75.1%	498	72.3%
nsimce	31	35.6%	46	7.6%	77	11.2%
Total	87	100.0%	602	100.0%	689	100.0%

Notes: To simplify the notation, subscripts are omitted, but each variable is centered on it respective cutoff and divided by their standard deviation.
Source: Authors' calculations.