Does school accountability pressure improve school quality?

Barbara Hanisch-Cerda

Submitted in partial fulfilment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2017

ABSTRACT

Does school accountability pressure improve school quality?

Barbara Hanisch-Cerda

This dissertation evaluates the impact of accountability pressure on a context where there is school choice. The Chilean context provides a unique opportunity for this purpose since it constitutes a system with school choice since the 1980s and since 2008 there is a policy that introduces a means-tested voucher that introduces incentives to schools to improve their performance. Under this policy, schools are classified based on students' test scores and other school factors (such as teacher evaluations, approval rates, retention rates), and are linked to punishments and rewards. I assess the impact of accountability pressure on outcomes that have consequences attached (high-stakes outcomes), and other outcomes that may reflect the quality of the school but do not have direct consequences attached (low-stakes outcomes). I use a fuzzy regression discontinuity to estimate the effect of receiving different school classifications on high-stakes test scores, low-stakes test scores, school behavioral responses, and student and teacher body composition. Estimates of the effects on 4^{th} and 8^{th} grade math, language and science are never significantly different from zero. There is also no evidence of parental or teacher response.

Table of contents

List of Tables

List of Figures

ACKNOWLEDGEMENTS

I would like to express my gratitude to my professors, colleagues, family and friends for all their support throughout all the years of the PhD.

Special thanks to Professor Henry Levin who has been my academic advisor since I joined Teachers College, being with me with the ups and downs of the PhD journey. I am indebted for his invaluable support, guidance and influential insights. Thank you for supporting me all these years, both personally and professionally as a researcher and teacher.

I also thank Professors Judith Scott-Clayton, Sarah Cohodes, Miguel Urquiola and Peter Bergman for your extensive and very thoughtful comments on my work on this dissertation. I am sure your advice will guide my research endeavors beyond this dissertation.

I am indebted to Violeta Arancibia. You guided me through my first steps into psychology and education. You have been my mentor in so many ways. Thank you for pushing me to pursue my PhD in economics and education.

I want to express my gratitude to my colleagues for fantastic moments together, both working and relaxing. Special thanks to my friends Siow Chin Ng, Yilin Pan, Sabine Zander, Martha Kluttig, JiHye Kim, with whom I laughed and cried and from whom I got unconditional friendship, encouragement and support in this long journey of the PhD.

I also would like to thank fellow researchers at CBCSE with whom I learned to apply most of the things I learned during the PhD: Fiona Hollands, Clive Belfield, Brooks Bowden, Ilja Cornelisz, Emma Garcia, Rob Shand, Ipek Bakir, Meridith Friedman, Henan Cheng, and Amritha Menon. Special thanks to Fiona Hollands for all your guidance, support and

determination in giving life to all the projects in which we worked together. Thank you for believing in me and my ideas. It was a pleasure to construct and deconstruct so many ideas, problems and solutions with you. I am thankful for the opportunity of working with you.

I am deeply grateful to my friends for all your unconditional support during these years: by far the happiest and hardest years of my life. Thank you for being there with me and my family. Thank you to Mary for your friendship, companionship, emotional -and practical- support. Thank you to Susan and Chip for receiving us in your home, and caring for us as part of your family. You have crossed my life as guardian angels.

I am grateful for the unconditional support of my parents and siblings. No words can express the gratitude, love and admiration I have for each one of you. Without your encouragement and love I would have not pursued the PhD.

Finally, and most importantly, I want to thank my husband Diego. Thank you for your eternal patience and support. Thank you for believing in this project. Thank you for sacrificing all those projects you had made commitments to in order to join me in this adventure. I am extremely fortunate to have you in my life, and for giving me the most adorable son, Simon.

I am grateful to CONICYT, Columbia University, Wolfensohn Foundation, and Teachers College for financial support.

DEDICATION

To Diego and Simon with all my love.

Does school accountability pressure improve school quality?

Chapter 1 - Introduction

It is believed that school accountability improves the quality of education. This belief underlies policies of school accountability, like "No Child Left Behind" in the U.S., which requires that states adopt single, statewide accountability systems. Like this, school accountability policies are flourishing throughout the world. But there is still a lot to learn about the impact of these types of policies.

The basic notion of school accountability policies is that attaching consequences (explicit or implicit, positive or negative) to students' performance will act as an incentive to school personnel and teachers to adopt actions that will improve students' performance.

Empirical evidence on the impact of school accountability has accumulated in the U.S. and England. There is a growing body of evidence that suggests school accountability improves students' achievement (Carnoy & Loeb, 2002; Hanushek & Raymond, 2005). Specifically, most of the studies that assess specific accountability programs focus on the impact of school accountability on performance measures associated with the incentive (high-stakes outcomes). Since the performance measures are just part of the educational goals of schools, other studies focus on the impact of the accountability policies on performance measures not associated with the incentive (low-stakes outcomes). The evidence shows that school accountability policies have a positive, but modest, impact on high-stakes outcomes (Jacob, 2005; Neal & Schanzenbach, 2010; Chiang, 2009; Rockoff & Turner, 2008; Figlio & Rouse, 2006). The evidence of the impact on low-stakes outcomes is mixed (Chiang, 2009; Jacob, 2005). The

evidence of the impact on long-run outcomes is too scant to draw conclusions (Deming, Cohodes, Jennings & Jencks, 2013).

The mechanisms through which school accountability operates to increase high-stakes tests scores can depend on (i) the behavioral response of the school, (ii) student body composition, and (iii) teacher body composition. The behavioral response of the school to increase high-stakes test scores can be both desirable and undesirable. Desired mechanisms are those that improve the educational goal[1] as well as the performance measure. Less desired mechanisms are behaviors that improve the performance measure at the expense of the educational goal. Among the desired mechanisms, there is evidence that schools increased instructional expenditures (Chiang, 2009; Craig, Imberman & Purdue, 2013; Rouse, Hannaway, Goldhaber & Figlio, 2013). Among the undesirable mechanisms, there is evidence that schools focus on marginal students at the expense of other students (Neal & Schanzenbach, 2010; Deming et al., 2013), schools narrow the curriculum (Jacob, 2005); shift instruction towards tested material (Jacob, 2007), manipulate the pool of test takers (Cullen & Reback, 2006; Jacob, 2005), increase special education placements (Jacob, 2005), and increase student retention in grades prior to the high-stakes grades (Jacob, 2005). Studies about the impact of accountability measures on distribution of students and teachers across schools are few. So far the evidence shows school accountability impacts student sorting among schools (Hart & Figlio, 2015) and affects the dynamics of the teacher labor market (Clotfelter, Ladd, Vigdor & Aliaga Diaz, 2004; Feng, Figlio & Sass, 2010).

[1] This assumes there is some sort of reasonable unanimous goal of schooling. However, attributing just one goal to schooling is quite reductionist. Schools serve several constituencies (e.g., parents, employees, state, teachers, students) all of which have their own set of goals; goals that may conflict with each other. In most instances, however, the goals of schooling are assumed to be the achievement of skills in subjects such as math, language and science measured by some standardized test (Levin, 1974).

In this dissertation I examine the first attempt of school accountability in Chile where schools are accountable to the government for their academic results, the law of adjusted vouchers, or SEP law (for its acronym in Spanish *Ley de Subvencion Escolar Preferencial*).This law offers extra funding for schools enrolling socioeconomically disadvantaged students, with the allocation of those extra funds conditional on schools' academic performance. Schools are classified into three performance categories ("in recovery", "emergent", and "autonomous"). This classification allows identifying schools that require more support to improve their quality. And, since Chilean parents are free to choose schools, this classification provides information for parents to choose the school for their offspring.

In accordance with the international literature, the empirical evidence on the impact of the introduction of the SEP law shows a positive but moderate impact on high-stakes outcomes, ranging between 0.08 and 0.2 standard deviations on a national standardized test (MINEDUC, 2012; Correa et al., 2013; Villarroel, 2012; Mizala & Torche, 2013; Neilson, 2013; Navarro-Palau, 2015). However, these studies have not disentangled the effect of the extra resources provided to schools from the effect of the school accountability component. There are no studies in Chile about the impact of the SEP law on either low-stakes outcomes or on long-run outcomes.

In contrast with current research of the SEP law, I do not limit my study to the impact of the introduction of the law as a black box on high-stakes outcomes. I try to study what incentives within this law affect the educational outcomes of the students. I specifically study the impact of accountability pressure on high- and low-stakes outcomes. Specifically I explore three questions about the ways in which school accountability affects students' educational outcomes:

(1) What is the impact of accountability pressure on high-stakes outcomes?

(2) What is the impact of accountability pressure on low-stakes outcomes?

(3) If school accountability increases incentivized or non-incentivized outcomes, what mechanisms drive those improvements?

I use a fuzzy regression discontinuity design that exploits the discontinuous structure of school classification in the Chilean accountability system. The key finding is that there is no consistent evidence that schools receiving different classifications affect high- or low- stakes outcomes of the students the year of the classification or a year after. Point estimates are often close to zero, however, never statistically different from zero. I perform relevant specification tests, and find that the lack of evidence of an impact holds throughout specification and robustness checks. I discuss several explanations of the findings.

This study contributes to the existing literature in two ways. First, school accountability policies have become very popular worldwide; however, there is little evidence on their impacts outside the U.S. and England. This dissertation provides evidence from Latin America. Second, this dissertation provides empirical evidence on the impact of school accountability on an educational system where there is school choice (Hart & Figlio, 2015; Mizala & Torche, 2013; Neilson, 2013). This paper also contributes specifically to the Chilean evidence about the impact of the SEP law in three ways. First, this study assesses the impact of accountability pressure on schools. Second, this study assesses the impact of accountability pressure on both high- and low-stakes outcomes. Third, this study assesses potential mechanisms that could drive the impact of accountability pressure on educational outcomes.

The reminder of the dissertation is structured as follows. In chapter 2 I provide a review of the literature of school accountability. In chapter 3 I review the Chilean accountability system and the empirical evidence accumulated so far. Chapter 4 presents the method for estimating the effects of school classification on educational outcomes and the data. In chapter 5 I present the findings. Chapter 6 summarizes and discusses the implications and limitations of this study.

Chapter 2 - School accountability literature

So far I have referred to 'incentives' and 'accountability' in a very loosely way. In order to move forward in the study of incentives and accountability I will stop for a moment to define these terms and their relationship. Then I review the theory behind school accountability and the empirical evidence accumulated so far.

2.1. Defining incentives and accountability

Incentives are rewards or benefits that promote certain actions or greater effort. In psychological literature, incentives are often mentioned along with the term 'motivation'. One simple approach to relate both concepts is that motivation represents the reasons for people's behavior or organizations' actions (Ryan & Deci, 1999). These reasons can be intrinsic to the activity, such as pleasure, feeling of accomplishment, etc. The reasons can also be extrinsic to the activity, i.e., that engagement in certain behavior is to obtain an outcome that is separable from the activity itself and therefore could also be obtained by doing other activities/behaviors. Extrinsic reasons can be rewards or sanctions if the behavior is or is not displayed, or the performance is or is not achieved satisfactorily. These extrinsic reasons, which can be explicit or implicit, are also called incentives.

In education, performance incentives can be described as:

> *"(…) rewards and punishments related to the achievement of specific outcomes. Incentives may be promised in advance, such as those written in teacher contracts, or given spontaneously, such as giving teacher a choice classroom assignment after a successful year. Incentives may also be monetary (such as year-end bonuses for bringing all students in a class up to grade level in reading) or implicit (such as the praise a teacher gets from colleagues for doing a particularly good job)" (Hanushek et al., 1994, p. xx-xxi).*

The way of understanding accountability policies is quite broad in terms of to whom they affect and what devices are set in place. For example, tests can have direct consequences for students serving as a powerful incentive for students to put greater effort into learning (OECD, 2013). Or students' assessments can have consequences for teachers' payments (Podgursky & Springer, 2007). Also students' assessments can be used by schools to allocate resources or to provide additional support for low performing schools. Some authors consider as an accountability policy any reporting of school performance by an external organization such as the government (Figlio & Ladd, 2015; Figlio & Loeb, 2011; Figlio & Ladd, 2008). Some others consider as accountability policies those measures that provide clear outcome thresholds the individuals or organizations are expected to meet with explicit consequences if those thresholds are met or not (National Research Council, 2011). Accountability policies can also include school choice policies, where the consequences are enacted by parents as they choose particular schools for their children and not others (Gronberg & Jansen, 2006; Ladd, 1996). Most authors do not explicitly favor one approach over another.

The common meaning seems to be that accountability policies are institutional devices that attach consequences (explicit or implicit, positive or negative) to measured student performance (academic outcomes). The purpose is to use performance information to increase students' outcomes. Such performance information is meant to act as an incentive for individuals and organizations to increase performance.

In summary, external incentives are the extrinsic reasons that drive behavior, and accountability and are the institutional devices that attach consequences to outcomes through the provision of information. The presumption is that such information will act as an incentive for individuals or organizations to improve those outcomes. Accountability provisions in education

constitute the policies that have been popular in the last decade. Incentives are the rationale behind such policies.

Accountability devices have been designed in several different ways. The common element that accountability policies in education promote is that their ***goal*** is to increase students' performance. There are features that shape them (O'Day, 2004). These features are:

Who is accountable? The target of reward or sanctions may differ between different audiences. For example, the target can be schools or private firms as organizations, or individuals such as school personnel, parents, or students. Who is accountable can also be found in a voluntary or mandatory situation. In a voluntary scheme, the constituency that is accountable gets into the relationship voluntarily searching for recognition or profits. In a mandatory scheme, schools or the constituency that is accountable is compelled to obey the accountability requirements in order to operate in the system.

To whom are they accountable? The constituency to which the target group of the reward or punishment is accountable also varies. It could be parents in the cases where parents choose the schools for their offspring. It can also be any sort of public organization that holds schools or teachers accountable for the results of their students.

For what are they accountable? The target group whose performance is affected can vary. It can be either all students of the school, a subgroup of students in the school, or a subgroup of the population regardless of what school they attend. Also the specific performance goal may differ. It can refer to increasing students' performance in reference to a baseline period, or to reach a basic standard. It can refer to specific subjects like math, reading, etc. There can also be specific goals for specific populations.

With what consequences? The rewards and sanctions also vary. It can be a monetary reward, or recognition or any other type of reward. The reward can be either explicit or implicit. Usually there is a combination of rewards, explicit and implicit. For example, there can be a monetary reward for a school for increasing the performance of their students, but there is also the reward of recognition in front of the community. Sanctions can also be explicit or implicit, and can be monetary or non-monetary.

There are several accountability devices with differing features that have been attempted over the years. There are devices where schools, teachers and students are accountable to either parents or public organizations. For example, schools are accountable to parents in a system of 'school choice', accountable to public organizations in 'performance-contracting', 'school accountability' and 'social impact bonds'.

Incentives and corresponding accountability devices are appealing as (i) they direct behavior towards specific goals, at least those goals that count towards receiving the rewards or avoiding the sanctions (Hannaway, 1996); (ii) they have the potential to align the behaviors of teachers and principals within public schools and districts with policy makers' and parents' broader goals (O'Day, 2004; Figlio & Loeb, 2011); (iii) they do not attempt to dictate which teaching methods will work, but encourage individuals to decide for themselves which route towards the goal is most appropriate for their specific circumstances (Hanushek et al., 1994; Hanushek, 1996); (iv) they reward teachers, schools or firms only for successful teaching (Garfinkel & Gramlich, 1973). In the long run incentives could lead to even more fundamental changes in existing educational systems. Those schools, teachers or firms that were successful in teaching would flourish and expand; those that failed would abandon teaching. Public organizations and parents would be given a chance to choose from competing sources of supply, choosing schools on the

basis of their outputs –the pupils who succeed- instead of inputs –the number of students, room space, and the like (Gramlich & Koshel, 1975).

The challenges with applying incentives to the management of education is (i) defining what are the goals that schools should seek; (ii) finding reliable measures of valuable schooling outcomes (not only short-run high-stakes measures, but long-run measures); (iii) isolating the impact of teachers and schools on student outcomes; (iv) defining what sort of incentives will work most effectively without also having undesirable side effects (Hanushek et al., 1994) like narrowing of the curricula, suffocating creativity, undermining student engagement (Woessmann, Ludemann, Schutz, & West, 2007), misbehavior, discouraging the pool of test-takers, unintended targeting of subgroup of students, and so on.

2.2. School accountability

There are two broad types of accountability systems that have the school as the unit of analysis[2]: school choice and school accountability (Hanushek et al., 1994; Gronberg & Jansen, 2006; Figlio & Loeb, 2011; Figlio & Ladd, 2008; Figlio & Ladd, 2015). The first type of accountability system operates by altering the structure of the school system (Hanushek et al., 1994) and without having the national or state governments prescribe what a school of quality is (although information systems may identify 'quality' dimensions). Schools are accountable to parents and students through a system of school choice usually with voucher-type programs or charter schools. Parental demand will supposedly indicate which schools are of high- and low-

[2] Levin (1974) describes four different ways of understanding what educational accountability is: as performance reporting, as a technical process, as a political process and as an institutional process. The way I approach what school accountability is on this dissertation is as a technical process. That is, I understand school accountability as a technical approach for evaluating the operations of the schools and improving the achievement of the goals of the schools. This approach assumes there is some unanimity on the goals of schooling.

quality, and their choice will reward high quality schools and punish lower quality schools expelling them from the market or making them less profitable. This type of accountability system is known as ***school choice*** (or 'voucher programs' or 'charter programs').

The second one, which is the one I focus on, requires setting performance standards schools have to meet. Supposedly these standards will reflect the expected outcomes of a school of quality. The main task is defining performance standards, defining appropriate performance measurement, identifying the desirable performance levels, and establishing an appropriate incentive scheme to reach those standards (Gronberg & Jansen, 2006; Figlio & Loeb, 2011; Figlio & Ladd, 2008; Hanushek, 1996). Good and bad performance of schools has explicit consequences, and those are evaluated and rewarded or sanctioned by the government or other public organizations. This accountability system is centralized in government agencies and operates within the structure of schools (Hanushek et al., 1994). This type of accountability system is known as ***school accountability*** (also known as 'new school accountability', 'test-based accountability', 'standards-based accountability', 'outcome-based accountability', 'check-up approach', 'consequential accountability', 'performance management systems'). Among this type of accountability are policies such as No Child Left Behind (NCLB)[3].

[3] For the most part, school accountability mechanisms are mandatory for schools. However, there are two variations where school accountability is not mandatory. One is ***performance-contracting*** where standards and performance thresholds are also required in order to enact contractual consequences. However, it differs from school accountability schemes as they are commonly known in that school accountability is usually mandatory for all schools operating in the system, whereas performance-contracting operates when there is a school or private firm voluntarily engaging in such contractual relationship. Such schools or private firms usually seek profits from such relationship (Peterson, 1974). A similar and more current type of financing mechanism associated with performance is ***social impact bond*** (also called 'pay-for-success bond', 'social benefit bond', 'development impact bond'). This is an outcome based contract where the public sector pays a proven service provider for the outcomes delivered. Private investors or donors fund the intervention. The total payment that returns to the investor is on the basis of improved long-run outcomes. If outcomes do not improve, then investors do not recover the investment (Social Finance, 2011). Investors also act as another constituency to which the service providers (which in the case of education could be schools) are accountable to. Social impact bonds differ from performance-contracting, in that social impact bonds focus on long term outcomes, whereas performance-contracting focus on short term outcomes.

2.3. Theory

Researchers have approached the study of school accountability from a variety of theoretical perspectives. The organizational learning and adaptation literature highlights the role of information on complex systems, and how information travels, interacts and is used to improve school quality (e.g., O'Day, 2004). Other experts have approached the issue from the educational change literature, emphasizing that educational policies should promote accountability devices that promote trust and collective responsibility rather than competition and control in order to improve educational performance (e.g., Sahlberg, 2010). The principal-agent perspective has focused on the need to reduce opportunistic behavior of the schools (e.g., Dixit, 2002; Ferris, 1992; Baker, 2002). I review the theory behind school accountability based mainly on the principal-agent perspective[4]. I do so using the example of a school district being the principal and schools being the agencies.

Whenever a firm hires an employee or a school district hires a school, a contractual relationship is stablished. The written contracts, however, are almost always incomplete, because it cannot specify all possible contingencies (in an ideal world, if it did specify all possible contingencies, then the contract would be prohibitively expensive).The firm has some set of goals and expectations about the behavior or outcomes from the employee that will help the firm attain those goals. The employee has other set of goals, and also has knowledge about what he or she can do for the firm. The firm, however, does not have complete knowledge about the personal goals of the employee, the skills of the employee, the effort he or she may put into the

[4] A 'principal', in the principal-agent relationship is mainly a 'boss' (a property owner if ownership is private, or the government agency if the ownership is public). I refer to 'principal' as this boss. If I need to refer to a school principal, I will refer to it as 'school principal'.

work and what he or she can attain in the future. There is a problem of asymmetric information and incomplete contracts.

The principal-agent framework focuses on how organizations build incentives into contracts given the goals of the organization, the limited information about the future effort workers will put into their work, the uncertain technology, the uncertainty about all possible situations and challenges the workers may face in the future, and the willingness of the worker of bearing risks of the uncertain outcomes (Driver, 2003). The principal-agent framework describes the relationship between a theoretical 'principal' (e.g., boss) and a theoretical 'agent' (e.g., employee), where the principal contracts an agent to act on his behalf (Ferris, 1992). In the educational system there are a series of principal-agent relationships; for example, when parents contract (with their votes) a school district to educate their children, or when school districts contract schools to educate the citizens (Ferris, 1992), or when 'school principals' hire teachers (Levin, 1980), or when superintendents hire 'school principals'[5] (Driver, 2003). This relationship between a principal and an agent, or specifically between the state or local governments and the schools, provides a rational for school accountability (Figlio & Loeb, 2011; Figlio & Ladd, 2008; Figlio & Ladd, 2015; Woessmann, et al., 2007).

Problems appear in the relationship between the principal (e.g. school district) and the agent (e.g. schools) when (i) the agent and the principal have different objectives and (ii) the principal has imperfect information to assess the behavior or performance of the agent (Ferris, 1992; Levin, 1980; Driver, 2003). Divergent objectives may appear if schools and school districts have

[5] To my knowledge, Driver's (2003) study of the relationship between 'school principal' superiors and 'school principals', is the only study that attempts to check whether the principal-agent theory in fact can be applied to the educational field. All other studies that make references to the principal-agent theory in the field of education seem to refer to the conceptual elements of the theory and not so much on how it is grounded in reality.

a different conception of educational performance, or if schools also consider other objectives that compete with educational performance such as caring about the employment conditions of the staff, offering music and athletics activities for students or using innovative pedagogical approaches (Ferris, 1992). Asymmetry of information (the agent having more information than the principal) may appear as school behavior and performance is hard and costly to monitor. Under these conditions three main types of problems appear: adverse selection problem[6], outcome verification problem[7], and moral hazard problem. I focus on the latter.

A moral hazard problem arises once the contract has been agreed and there is asymmetry of information between the agent and the principal. The agent, realizing the principal's lack of information, pursues the agent's objectives at the expense of the principal's (Ferris, 1992; Fernandez, 2009). For example, if school districts have difficulty monitoring the behavior of schools, then schools ('school principals' or teachers) might behave in a manner contrary to the interests of the school district, that is, they may not put the necessary effort in the assigned task (Levin, 1980; Driver, 2003).

Moral hazard problems can be addressed using performance-based contracts to ensure the accountability of the agent (Ferris, 1992) [8]. By monitoring the behavior or performance of the

[6] Adverse selection arises when the principal lacks information about the skills of the agents, which may lead to inappropriate decisions about whether or not a certain agent should be hired (Ferris, 1992). This problem appears before there is a contractual relationship between a principal and an agent (Dixit, 2002).

[7] The outcome verification problem arises when the agent can observe some outcome better than the principal (Dixit, 2002).

[8] The feasibility of performance-based contracts as a solution for a moral hazard problem depends on: (i) the cost of monitoring the behavior or performance of the agent, (ii) the correlation between the processes of production and the performance of the agent, and (iii) how much risk the agents are willing to undertake (Ferris, 1992). If the cost of monitoring the performance of the agent is less costly than the value of producing the expected outcomes, then the principal may just monitor the performance of the agent himself (Ferris, 1992; Fernandez, 2009; Holmstrom & Milgrom, 1994). Indeed, monitoring expenses can be considerable and they may outweigh any benefits derived from monitoring (Fernandez, 2009). If the cost of monitoring the performance of the agent and the correlation between the production processes (behavior) and the performance is high, then the principal may just monitor the behavior of

agent, the principal can detect opportunistic behavior. By attaching positive consequences to strong performance and negative consequences to weak performance, the principal provides incentives for the agent to align the agent's behavior with his own goals[9]. A performance-based contract in education could be a contract that determines school funding conditional on school performance, perhaps not tying all the budget of the school to performance, but part of it. Another possibility is that the school district removes decentralized-decision making power if schools do not improve their academic performance; or offers the relaxation of some regulations if the schools improve their academic performance (Ferris, 1992). More effective monitoring of educators could result in improved student outcomes (Figlio & Loeb, 2011; Figlio & Ladd, 2008; Woessmann et al., 2007).

Assumptions. The theory of action behind accountability policies posits that the threat of sanctions and possibilities of rewards will incentivize schools to align their behaviors with stakeholder's expectations, and in turn, this alignment will improve students' achievement

the agent and not the performance. If the correlation between the production processes and the performance is low, then the principal will be interested in monitoring the performance of the agent and not the processes of production. This is particularly relevant in education as the educational performance of students depends on what happens in the school as well as on what happens at home. Now, if the agent is risk-averse he may not be willing to accept a contract where there is low correlation between the processes of production and the outcomes, since the agent is the one who assumes all the risks of the contract.

[9]There is some sort of consensus that punishments and rewards are differently valued by subjects. Negative outcomes have larger value than positive outcomes (Baumeister, Bratslavsky, Finkenauer & Vohs, 2001). Individuals are more upset when they lose '*x*' money than more happy to win '*x*' money (Kahneman & Tversky, 1984). There is also consensus that punishments and rewards have an effect on behavior; however, whether punishment and rewards generate qualitatively different behavioral effects is still a matter of study. Thorndike's (1927) perspective was that rewards increase behavior frequency, whereas punishment decreased behavior frequency by the same magnitude it was increased by the reward. In this perspective, the fear of losing a reward is equivalent to being punished. Others view rewards and punishments as qualitatively distinct (Yechiam & Hochman, 2013). Current studies suggest rewards and punishments are qualitatively different. Kubanek, Snyder and Abrams (2015) performed a simple experiment were they varied the magnitude for the reward or penalty after individuals made a choice. As expected, they found a reward led the individual to repeat the choice of the previous trial, and a punishment led the individual to avoid the choice of the previous trial. As they varied the magnitude of rewards and punishments, the authors found that as the magnitude of the reward increased, so did the repetition rate of the prior trial. However, regardless of the magnitude of the punishment, the avoidance rate of the prior trial was flat. This suggests rewards and punishments are different factors affecting behavior.

(O'Day, 2004; Figlio & Loeb, 2011). This theory of action has several underlying assumptions. These assumptions are that (Driver, 2003; Fuhrman, 2004; Baker & Linn, 2004):

1. Performance, or student achievement, is the key value or goal of schooling, and constructing accountability around performance focuses attention on it[10].
2. Performance is accurately and authentically measured by the assessment instruments in use. This also means that the results reported are accurate and reliable.
3. States correctly interpret information about students' achievement and know how to use that information to improve achievement. E.g. to implement sanctions and rewards or to offer technical assistance to schools.
4. School personnel correctly interpret information about students' achievement and know how to use that information to improve achievement.
5. Consequences, or stakes, motivate school personnel, i.e., school personnel and teachers see sanctions as a threat and rewards as desirable.
6. School personnel know alternative actions to improve the situation[11]. Cognizant individuals and team members possess the requisite knowledge to apply alternative methods. The selected action is adequately implemented.
7. The action selected will improve instruction and higher levels of performance will result.
8. Parents also correctly interpret information about students' achievement and know how to use that information to improve achievement. E.g. for voicing dissatisfaction or voting, or if

[10] Of course, this key value or goal will almost always be an incomplete measure of output, especially in a multi-product firm.

[11] Assumption 6 brings about immediately assumption 9. The assumption that school personnel know alternative actions supposes they are desirable actions that aim to improve learning and not only test scores. Opportunistic behavior that may increase test scores and not learning, such as teaching the test, are among those "unfortunate unintended consequences" supposed to be minimal (mentioned in assumption 9).

the context is a system of school choice, then parents would use the information to choose schools.

9. Unfortunate unintended consequences are minimal.

   These assumptions could be tested, as in the exploratory analysis of Driver (2003).

Moderators. The solution above suggests accountability, by aligning the agents' behavior with the goals of the principal, should have a positive effect on the performance of the agency. Such relationship as depicted above is quite simplistic. The impact of accountability may be moderated by other factors that influence the performance of the agent (Fernandez, 2009) complicating monitoring. Some of these factors are: (i) the amount of tasks undertaken by the agent, and whether those tasks are complementary or substitutes (Holmstrom & Milgrom, 1994; Fernandez, 2009; Driver, 2003; Dixit, 2002; Laffont & Martimort, 2002; Baker, 2002), (ii) the uncertainty or complexity of the tasks undertaken by the agent (Levin, 1980; Hannaway, 1996; Dixit, 2002; National Research Council, 2011; Neal & Schanzenbach, 2010) and whether the relationship develops over a period of time or one time only (Kane & Staiger, 2002), (iii) whether there is one or multiple principals (Driver, 2003; Dixit, 2002), (iv) whether the agent is a profit or nonprofit organization (Gramlich & Koshel, 1975; Peterson, 1974; Fernandez, 2009), and (v) whether there is competition among agents (Carpenter-Hoffman, Hall & Sumner, 1975; Dixit, 2002).

Design of the accountability scheme. The impact that accountability could have on performance of the schools not only depends on the assumptions of the theory of action and the moderators depicted above, but also depends on the minutiae of the design of the accountability scheme (or specifically of the performance-based contracts). The performance-based contracts may or may not provide an incentive for the schools to improve performance depending on how the contracts

are designed (National Research Council, 2011). It is not the purpose of this paper to delve into these issues (for that see Figlio & Loeb, 2011), but some design issues that experience has shown to be important in the educational context are: (i) how close is the performance measure to the intended goal (Ferris, 1992; Baker, 2002; Kane & Staiger, 2002; Richards & Sheu, 1992; National Research Council, 2011), (ii) what performance indicators are being used and whether they are means or proficiency levels (National Research Council, 2011; Figlio & Ladd, 2015), and whether it is one measure or multiple measures (Bush, Hough & Kirst, 2017), (iii) what time span is used to measure performance and whether it is a static measure or a measure of growth (Linn, 2004; Kane & Staiger, 2002; Figlio & Ladd, 2015; Figlio & Ladd, 2008; Figlio & Loeb, 2011; National Research Council, 2011), (iv) who is in the tested group (Figlio & Ladd, 2008; Figlio & Loeb, 2011; National Research Council, 2011; Figlio & Ladd, 2015), and (v) how performance relates to the rewards or sanctions (Dixit, 2002).

2.4. Empirical evidence

The main challenges of assessing the impact of school accountability are three. First, usually school accountability is part of a set of standards-based reforms. Disentangling the effect can be rather convoluted. Second, school accountability policies are usually set in place for the whole system making it hard to find appropriate control groups to identify what would have happened if no accountability policies were enacted (Figlio & Ladd, 2011). Third, school accountability systems may have impact on students' achievement, not necessarily by the improvement of schools' and teachers' practices, but through opportunistic behavior of schools (e.g., cheating, narrowing of the curriculum, and reclassification of students into non-tested groups). Understanding which mechanism is in place is relevant to inform policy.

Impact evaluations of school accountability can be broadly grouped into two categories (Figlio & Loeb, 2011; Lee, 2008). The first group is all those studies that compare different school accountability systems such as cross-national studies or cross-state studies (e.g., Carnoy & Loeb, 2002; Hanushek & Raymond, 2005). These studies usually have the states or nations as the primary unit of analysis. They use the variation across nations or states to assess the impact of the policy (Figlio & Ladd, 2011; Lee, 2008; Woessmann et al., 2007). These studies use independent international or national low-stakes tests to compare the results across nations or states. The advantage of these studies is that their conclusions are not too idiosyncratic and are generalizable to school accountability policies in a broad sense (Figlio & Loeb, 2011; Figlio & Ladd, 2008; Figlio & Ladd, 2015). Such studies allow responding to the question of whether school accountability systems have an impact on students' achievement. The disadvantage is that the conclusions do not refer to any specific type of policy design (Figlio & Loeb, 2011; Figlio & Ladd, 2008; Lee, 2008). Another disadvantage is that usually school accountability is part of a bundle of other policies (e.g. increased funding for schools), and if the effects of a single policy in a state is not disentangled from the effect of the other policy adopted at the same time, then the effects may be biased. Most of the studies on this group use a purely empirical approach without trying to understand the mechanisms through which accountability might have affected student outcomes (Lee, 2008).

The second group is all the evaluations that refer to a specific school accountability system, that includes nation-, district- or state-specific systems (e.g., Jacob 2005; Neal & Schanzenbach, 2010). The units of analysis of these studies are the schools or students within the system. These studies use the variation (i) across groups of students, (ii) timing of the policies or (iii) diverse

intensity (pressure) of the accountability policies on some schools[12] to assess the impact of the policy. The advantage of this type of study is that they may offer more clues about how is it that school accountability affects student's achievement. However, the disadvantage of these evaluations is that conclusions may not be generalizable to school accountability systems that have other types of design.

My interest is in studies from the second category, those evaluations of a specific system, as I am interested in contrasting this evidence with a well-defined accountability policy set in place in Chile.

Studies to determine the impact of district or state specific school accountability systems have focused on its impact on high-stakes outcomes (i.e., those performance measures that are attached to the incentive) and low-stakes outcomes. Fewer studies have focused on long-run outcomes. Studies that try to identify the actual mechanisms through which school accountability operates have focused mainly on the behavioral response of the schools. However, there are other potential mechanisms in place that can affect the impact of school accountability, such as the movement of students and teachers across schools. Fewer studies have focused on these mechanisms.

What effects do school accountability programs have on high-stakes outcomes?

[12] Studies of the effect of accountability pressure address the notion of 'pressure' in two different ways. Some studies consider the expected school classification as the treatment (e.g., Craig, Imberman & Purdue, 2013; Deming et al., 2015). This expectation is predicted given the rules of school classification. Other studies consider the actual school classification as the treatment (e.g., Figlio & Rouse, 2006; Rockoff & Turner, 2008). The first group of studies offers an overall measure of whether the pressure of that particular aspect of the policy is good or not. The second group of studies only assesses the impact on the treated.

Studying the impact on high-stakes tests is relevant for school accountability policies as it reflects whether schools are responsive to the policy. It may not be the best way of assessing the overall impact of school accountability, because of score inflation (Koretz, 2005; Jacob, 2007; National Research Council, 2011). Nonetheless, it is a relevant starting point. The findings of the impact of school accountability on students' high-stakes test scores are overall positive, but modest (see Table 1).

Studying test trends after the implementation of an accountability system in South Carolina, Richards and Sheu (1992) found modest improvements in student achievement, with larger achievements in low socioeconomic (SES) schools. Comparing pre-post test scores in Texas, Klein and colleagues (2000) found large changes in scores. This seemingly miraculous improvement was referred to, by some, as 'the Texas miracle' (Klein, Hamilton, McCaffrey & Stecher, 2000). However, none of these studies allow us to make causal inferences of the impact of school accountability on educational outcomes. It is not clear whether those trends could reflect other policies at the same time of the school accountability program or any other unobserved time-varying factors at the state or national level, including teaching to the test.

Studies that attempt to account for unobserved time-varying factors use differences-in differences to estimate the impact of the introduction of a school accountability policy or changes in the policy. Jacob (2005) studied the impact of the introduction of an accountability system in Chicago in 1996. Compared with other large urban districts, he found math and reading increased after the introduction of the policy. This finding is consistent with the findings of Neal and Schanzenbach (2010) also for the accountability system in Chicago. The latter authors went even further, and instead of focusing only on the average change in scores, these authors focused on the impact on different deciles of the prior achievement distribution. They

found that the increase in reading and math could be explained by increases in achievement of students in the middle of the achievement distribution. These same findings appear to be attributable to the introduction of the accountability NCLB policy introduced into Chicago and the Nation in 2002. This uneven distribution of changes in achievement implies accountability policies may only be beneficial for students within reach of the goal of proficiency, not those above or below.

Another group of research studies exploits the diverse intensity (pressure) of the accountability policies on some schools to assess the causal impact of the policy. These studies are more limited in scope as they do not estimate the systemic effect of the accountability system, but the effect of the pressure of facing the consequences of receiving a certain evaluation grade or classification. Studying the impact of such pressure on Florida, Figlio and Rouse (2006) found that schools graded with the lowest possible grade ('F') slightly increased their results in comparison to those schools that received any other grade, a similar finding to Chiang (2009). However, Chiang also found that the results do not seem to persist. Allen and Burgess (2012) performed a similar study in the English inspectorate system[13]. They found that schools that just failed the inspection improved their test scores, similar to Hussain's (2015) findings. However, contrary to Chiang's findings, Allen and Burgess found schools increased their outcomes as years passed, up to four years after the school failed the inspection. Rockoff and Turner (2008) studied the impact of schools receiving different grades on academic outcomes of elementary and middle schools in New York City. They found that receiving a low grade increased slightly math and reading achievement, and no evidence that receiving a higher grade impacted

[13] The English inspectorate system differs from the pure test-based accountability in that schools are not only accountable for their test scores, but also for a subjective evaluation of the school and classrooms performed by the Inspectors.

educational outcomes. Deming et al. (2013) studied the Texas accountability system. They used school fixed effects to compare students within the same school but across cohorts that face different degrees of accountability pressure. This allowed them to account for differences across schools in unobserved factors. They found that high schools that face the pressure of being rated low performing increased student achievement on high-stakes tests. The impact was larger on low achieving students.

When the achievement distribution of the students is accounted for, the impact of school accountability on systems based on proficiency seems to be higher on the middle of the achievement distribution at the expense of the left tail of the distribution. Neal and Schanzenbach (2010) found school accountability in Chicago increased the achievement of students in the middle of the achievement distribution. The authors did not find evidence suggesting that students in the extreme of the distribution benefited by the introduction of the accountability program. Consistent with this evidence, Deming et al. (2013) found in Texas that high schools that face the threat of being rated as low-performing schools increased students' achievement on high stakes exams especially for students who failed the year prior to the test. Furthermore, the authors found some evidence that schools that expected to be ranked as high performing showed negative effects on students who failed the exam the previous year. A similar story to the one found in Texas was found in the English inspectorate system. The evidence shows that low-ability students are the ones that benefit the most from a failing inspection (Hussain, 2015). The author suggests it is the role of the inspectors that makes the difference with purely test-based accountability.

The effect of school accountability not only seems heterogeneous for students with different prior achievement, but also for schools with different levels of quality. According to Jacob

(2005), students in low performing schools seem to have fared considerably better in high-stakes tests under the policy than comparable peers in higher-performing schools. This evidence is consistent with the one found in Texas by Deming et al. (2013), where schools graded as low-performing seemed to be more responsive to the incentives as they increased the number of students passing their high-stakes tests on time.

The effect of school accountability pressure seems to be stronger in a context where there is a big share of high-performing schools in contrast to a context where there are few high-performing schools. A recent study of Weiner, Donaldson and Dougherty (2016) focuses on the impact of accountability pressure on schools that just missed the high-performance classification benchmark in Rhode Island. The authors found that schools that just missed the high-performing benchmark, and are among several schools classified as high-performing, increase the academic outcomes of the students. This did not happen if the school was not surrounded by high-performing schools.

In sum, overall the findings show positive but mild effects of school accountability on high-stakes tests. Effect sizes of the impact on average scores range from 0.04 to 0.33. The evidence also supports there are differential effects on students in different positions in the achievement distribution, differential impact on schools of different quality, and different impact on schools with different levels of competition.

What effects do school accountability programs have on low-stakes outcomes?

A way to study whether the school accountability policy has affected educational goals beyond the performance measure attached to the incentive is to assess the impact on tests or grades for which there are no consequences attached in the incentive scheme (low-stakes test scores, or

low-stakes grades). This is the recommendation of the National Research Council (2011). The findings show mixed results (see Table 2).

Studying test score trends before and after the introduction of an accountability system in Texas, Klein and colleagues (2000) found a small increase in low-stakes test scores and an increasing racial gap. Also in Texas, Deming and colleagues (2013) found that the threat of being classified as low performing positively impacted graduation rates and math credits taken in high school; however the possibility of being recognized for good performance had a negative impact on high school graduation rates. Figlio and Rouse (2006) found that school accountability pressure on low-performing schools in Florida impacted their low-stakes results positively and statistically significant, but considerably less than for high-stakes tests. Also in Florida, Chiang (2009) found that the impact of sanction threats on low-stakes tests was negligible for both math and reading. Jacob (2005) found the introduction of the accountability system in Chicago had a positive, but mild effect, on 8th graders, a negative effect on 3th graders and no effect on 6th graders.

What effects do school accountability programs have on long-run outcomes?

The purpose of school accountability is to increase the educational achievement on the premise that such gains will lead to long-run improvements in educational attainment and earnings. There is very little evidence about this impact (see Table 3).

Deming et al. (2013) assessed the impact of accountability pressure in Texas high schools on long-term outcomes such as postsecondary attainment and earnings. Using school fixed effects the authors compared students within the same school, but across cohorts that faced different degrees of accountability pressure. The authors found that high schools respond to the

probability of being rated as low performing schools by increasing students' achievement on high-stakes exams. This effect appeared also on college attendance, completion of four-year degree programs, and 25 year olds' earnings. These effects were higher for low achieving students in low performing schools. The authors found there was no overall impact on high performing schools (those receiving higher accountability rating). Low achieving students in high performing schools were found to have even negative long term effects.

This evidence supports the idea that some of the responses of schools are reflected in real learning, and not only on the performance measures.

What mechanisms seem to drive the impact of school accountability on educational outcomes?

Studies about the mechanisms of policies may help us interpret the results of policy evaluations (Ludwig, Kling & Mullainathan, 2011). The mechanisms through which school accountability seems to operate to increase high-stakes tests scores depend on (i) the school behavioral response, (ii) the student body composition, and (iii) the teacher body composition.

School behavioral response. School behavior to increase high-stakes test scores can be both desired and not desired. Desired mechanisms are those that improve the underlying educational goal as well as the performance measure. These could be due to positive behaviors of principals, teachers, parents and students. Less desired mechanisms are opportunistic behaviors that improve the performance measure at the expense of the educational goal. To my knowledge there are no studies that assess the impact of the changes in the school practices due to the accountability policies on students' achievement, although the study of Rouse and colleagues (2013) is a first approximation. The existent studies mainly focus on the impact of the school

accountability policies on changes in the school practices (see Table 4). Their results provide some clues about the potential mechanisms that drive the impact of school accountability on educational outcomes.

Among desired mechanisms, there is evidence that schools had increased instructional expenditures[14] (Chiang, 2009; Craig, Imberman & Purdue, 2013; Rouse et al., 2013), and decreased principal control (Rouse et al., 2013). However, it is not clear that budgetary augmentation remained years after the schools suffered the shock of being evaluated as worse than expected (Craig, Imberman & Purdue, 2013).

Among the none desired mechanisms, there is evidence that schools focused on marginal students at the expense of students in the tails of the achievement distribution (Neal & Schanzenbach, 2010; Deming et al., 2013), schools narrowed the curriculum (Jacob, 2005), schools focused on the skills tested and testing formats (Jacob, 2007); manipulated the pool of test takers (Cullen & Reback, 2006; Jacob, 2005), increased special education placements (Jacob, 2005), and increased student retention in grades prior to the high-stakes grades (Jacob, 2005).

In sum, the evidence suggests the schools are responsive to the incentives as they do take action to increase high-stakes results. The way this is accomplished seems to range through a large set of mechanisms, both desired and not desired. Unfortunately, none of the studies that found positive effects on low-stakes tests studied potential mechanisms that may explain the

[14] A study by Jackson, Johnson and Persico (2015) presents evidence that expected increases in school spending due to school financial reforms has effects on long term outcomes such as increased years of schooling, higher wages, and a lower percentage of adult poverty.

effect. Therefore, we cannot tell whether any of the mechanisms that were found affecting high-stakes test may also affect low-stakes results.

Student body composition. The student body composition can also alter the educational outcomes of the school. Evidence about the impact of school accountability on student sorting between schools is scant (see Table 5). The information about the quality of schools may affect the way parents choose the schools for their children. For instance, high achieving students in low achieving schools may decide to migrate to high achieving schools. Or low achieving students in high achieving schools may somehow be pushed away from their school. Or perhaps low achieving schools could make more efforts to attract high achieving students. There are several possible ways in which the movement from one school to the other could occur. The sorting may be related not only to the prior achievement of schools and students, but to the access to information of school quality. Parents from high SES may have more information about the quality of schools, and therefore be responsive to the information. Low SES parents may not have much access. Or perhaps low SES parents are more responsive to the extra information provided about school quality (high SES parents may already have had information about the quality of schools).

Hart and Figlio (2015) studied whether the introduction of the accountability system in Florida was associated with changes in the composition of students across schools that received different grades. The evidence suggests more educated mothers responded to the extra provision of information to make a decision when choosing the school of their children. The effect was stronger when there were alternative schools nearby, and especially if the alternative schools received low grades.

The limited amount of existing evidence suggests school accountability impacts students sorting between schools, favoring high SES students (Hart & Figlio, 2015). This evidence suggests that school accountability could be enhancing stratification rather than reducing it. More research is needed in this area.

Teacher body composition. The teacher body composition can also affect the educational results of the schools after an accountability system has been introduced (see Table 6). Schools labeled as low performing may find it challenging to retain and hire teachers of good quality. High quality teachers in low performing schools may migrate into high performing schools as these may require putting less effort for teaching, and may provide benefits from more autonomy and affiliation with a successful school. Some schools may have more capacity to replace low performing teachers through high competition for each place. Although increased movement is not wrong per se, it can be bad if the least effective teachers end up in low performing schools.

Clotfelter et al. (2004) studied how the introduction of the *ABC* accountability program in North Carolina affected the rate at which teachers left low performing elementary schools. Looking at two cohorts of teachers (pre- and post- program) and comparing teachers in low performing schools with all the other schools, the authors found that teachers in low performing schools had a higher probability of departure. Feng, Figlio and Sass (2010) studied the effect of a change in the grading system in Florida's accountability system in 2002 on teacher's decision to stay or leave a school. The evidence suggests teachers were more likely to leave schools that had received lower grades than they had before, especially if the new grade was the lowest in the grading scale. Teachers in schools that were upwardly graded were less likely to leave their schools. The authors also assessed the quality of the teachers leaving and staying in the schools. They found that schools experiencing precipitous declines in rating had high quality teachers

leaving schools, and that the quality of teachers staying in these schools improved after the accountability policy was set in place. Teacher quality was measured as the value added of students' test scores. This sets some uncertainty in evaluating the findings, as we know from the literature that students' test scores, in an accountability setting, can be improved by undesirable mechanisms, and this issue is not addressed. In this study, a cheating teacher can easily be confused with a good teacher, as both appear to improve the performance measures of the students.

Overall, the evidence suggests there is increased attrition in low performing schools (Clotfelter, et al., 2004; Feng, Figlio & Sass, 2010). In addition, there is evidence that teachers tend to leave from schools with high poverty rates (Clotfelter, Ladd, Vigdor & Wheeler, 2006). Put together, this evidence could suggest low performing schools with high-poverty rates may be in a series disadvantage when it comes to attracting good teachers.

2.5. Conclusion

The school accountability programs help focusing attention on students' outcomes. Monitoring has provided information about the students' outcomes. This information can be useful in identifying students and schools that are most in need, and provide them with assistance.

The empirical evidence of school accountability programs shows students' outcomes have mildly increased high-stakes outcomes. Effect sizes range from 0.04 to 0.33. The effects on high-stakes outcomes seem to be heterogeneous for the different levels of prior achievement of the students, with the ones in the middle of the achievement distribution being the most benefited. The effect in high-stakes outcomes is also heterogeneous for the level of achievement of the school, with the schools that perform worst being the ones that most improve their results, most

likely because they are the ones that fear the largest sanctions. The empirical evidence on the impact of school accountability on low-stakes outcomes is not conclusive. These suggest school accountability may not be impacting true learning[15].

Several experts have assessed the mechanisms that drive the effect on high-stakes outcomes. Most of the mechanisms studied refer to changes in the behavioral response of the school. The evidence so far shows that increases in high-stakes outcomes may be due to such non desirable behavior of the schools as narrowing of the curriculum, manipulation of the pool of test-takers, increased student retention, increase number of special education placements, and focus on marginal students. Only two studies found schools increased instructional expenditures. Fewer studies have studied teacher and student mobility between schools as an effect of the school accountability programs.

Research is still required to address, at least, five main concerns about school accountability. First, what is the impact of school accountability on long-run outcomes? Some research has been done, but the accumulated evidence is not enough to yield generalizable conclusions. Second, how have accountability policies impacted student sorting? We cannot be promoting accountability policies as improving the quality of education without knowing if this is achieved at the expense of the less advantaged students. Third, what design is best to improve educational outcomes? It seems that school accountability policies are here to stay. We need to know what are the lessons learned about their design. Fourth, how much do the school accountability

[15] Similar research questions have driven research on performance-based contracts on higher education. The essential conclusion is similar to what is found on K-12 systems: It is not clear performance-based funding policies have improved college performance (Hillman, 2016).

programs cost in comparison to the acquired benefits. Are the benefits worth the costs? Fifth, does the effect of school accountability policies vary in contexts where there is school choice?

In the next chapter I describe a new school accountability policy in Chile and explain some of the empirical evidence of its impact accumulated so far.

Chapter 3 - School accountability in Chile

3.1. Historical overview

In 1980 Chile adopted a reform that transformed the system into a marketplace where there is freedom for parents to choose schools for their children, and where there is freedom for stablishing schools allowing any entity to create a school if complying with basic rules. The assumptions of the reform were that (i) competition between schools increases overall productivity, (ii) parental choice (with a demand-side subsidy) allows social mobility and ensures universal compulsory schooling, and (iii) private schools can be more efficient than public schools. The main transformations comprised financing and schools management. From 1980 until 2008 parents received a flat voucher, adjusted for school level, which allowed them to exercise their freedom to choose a school for their offspring. This reform also embraced a decentralized system, where public schools were no longer managed by the Ministry of education (MINEDUC), but by each one of the 238 municipalities.

This two-pronged reform led to changes in both the supply and demand side of schooling with intended and unintended consequences. It was expected that market forces will align the incentives of school administrators and teachers with parental demand, thus improving quality. The entrance of private enterprises in the competition for students was supposed to improve the efficiency of resources. The possibility of low income students choosing better schools than the ones near their homes was supposed to increase equity. However, some of the features that characterized the Chilean educational system by the year 2000 were parental freedom to choose, but with high levels of segregation, unequal distribution of expenditures per pupil and high outcome inequality. It seems the “market” was not enough to provide sufficient accountability among providers. As Deming and Figlio (2016) point out, in the education industry consumers

have limited power to ensure the quality of the services demanded. This limitation comes from geographical jurisdiction constraints, public provision of education is widespread, and the fact that the entry or exit of schools to the market is limited.

Since the 1990s, several programs have been established that attempt to improve the quality of the schools, new programs targeting low performing schools, providing parents with more information, or providing schools with incentives to improve. An example of a very well-known school intervention in Chile was the program called P900 which provided support for the lowest performing schools (Peirano & Vargas, 2005). An example of a policy that provided information to parents was implemented in 2010. Parents were provided with a map with schools colored according to their academic outcomes as in a traffic light. The academic outcomes reported were raw test scores without any adjustment per background or enrollment rates of the school. This intervention did not last long due to strong opposition of educational practitioners (Allende, 2012). Another example is the SNED program, which is a program of collective incentives for teachers if the school where they work achieves good results in comparison to schools with similar geographical and socioeconomic background (Mizala & Urquiola, 2013). It is in this context full of attempts of interventions to improve the quality of the schools that the SEP law is introduced.

3.2. SEP law

In 2008, a law of adjusted vouchers and accountability policy has been set in place: SEP law (Law 20,248)[16]. This law explicitly attempts to "improve quality of education of subsidized

[16] The SEP law (20,248) was enacted on January 25th of 2008. This law was regulated with the decree 235 from April 2008. This decree was later adjusted by the decree 293 from August 2009. The SEP law was later adjusted by the law 20,501 from February 2011, and by the law 20,550 from August 2011. I will explain these adjustments as I explain the different features of the law.

schools". For that purpose, the law mandates two lines of action. First, it adjusts school vouchers increasing its value for students from economically disadvantaged families (called 'priority students') up to 60% of the original value and adding a subsidy per concentration of priority students in a school. Second, it commit schools to improve school quality making all sources of public funding conditional on academic improvement (4 year span) and conditioning the management of the resources of the adjusted voucher to the academic results of the school.

This law is very relevant for Chile as it not only offers extra resources to schools, but it is the first time vouchers are given conditional on academic outcomes. Before the SEP Law, consequences from deficient outcomes were expected to be reflected in a lowered demand from parents. The SEP Law is the first regulation that attaches outcomes to consequences from the central government in Chile.

This law offers several incentives to schools. First, there is the incentive for subsidized schools to enroll more students from disadvantaged backgrounds. This is expected to increase social mobility and decrease inequalities, as students from low socioeconomic backgrounds now have more options to choose schools that in the past they could not pay for with additional fees. Second, there is the incentive for schools to improve their services. All schools that adopt the law are provided with significantly more resources. Such resources are intended for specific educational processes, such as improvement of curriculum, school climate, leadership and management of educational resources. This is expected to increase the educational outcomes of the schools. Third, schools classified as low performing have an incentive to perform better. Low performing schools face the threat of sanctions, and if they improve their performance they have possibilities of more autonomy to allocate the resources provided by the adjusted voucher.

With more detail, the features of the policy are:

Adjusted voucher. Until 2008, the per capita funding in Chilean schools was through a flat voucher (base voucher), differentiated by school level, regardless of the profile of the students. The SEP law increases vertical equity by giving an extra per-student subsidy for priority students who are enrolled in public or private voucher schools, and a subsidy per concentration of priority students. These students are identified by MINEDUC every year[17]. The group of priority students covers up to 40% of the poorest students from primary and secondary[18] education. The value of the adjusted voucher varies per grade level (see Figure 1). It adds up to 60% on top of the base voucher[19]. The value per concentration of priority student adds up to 10.5%[20].

Schools can allocate the resources of the adjusted voucher in four different areas: (i) curriculum management (improvement of pedagogical skills, extra help for students with special needs, class size reduction, more teachers and assistants, field trips, etc.); (ii) leadership (training for school management team, strengthening of teacher council, school participation of people from the cultural, scientific and local or national governance spheres, strengthening of the school-community relationship, etc.); (iii) school climate (support for students and families with

[17] To identify priority students, MINEDUC uses four criteria: (i) whether the family is in the program *Chile Solidario*: (ii) if not in *Chile Solidario*, families from the lowest third of socioeconomic background as defined by the characterization instrument in place; (ii) if not in any of the above, families in the range A of the *Fondo Nacional de Salud*; (iv) if not in any of the above, families can be selected by their level of income, schooling of mother (or father), rurality of the family, and degree of poverty of the municipality where the student lives, according to the regulation in place.

[18] Secondary education was incorporated to the law in year 2011 (Law 20,550).

[19] The law 20,248 from 2008 mandates an adjusted voucher that is up to 49% above the base value of the voucher. The law 20,550 from 2011 amends the value of the adjusted voucher increasing it up to a 58.7% above the base value of the voucher (PK-4th grade: 58.74%; 5th-6th grade: 39.02%; 7th -8th grade: 19.66%; and 9th-12th grade: 16.52%).

[20] The law 20,248 from 2008 mandates a voucher per concentration of priority students that is up to 8.7% above the base value of the voucher. The law 20,501 from 2011 amends the value of the voucher per concentration, increasing it up to 10.5% above the base value of the voucher.

psychologists and social workers, strengthening of school council, school climate improvement, etc.); and (iv) resource management (teacher training, student evaluations, teachers and management team incentives for performance, improvement of school library, etc.). Schools need to report to MINEDUC how they are allocating the extra resources provided. However, the extra resources need not be allocated exclusively to priority students. Also, the schools have no restriction in having the new SEP funds substitute for other funds that would have been spent on the educational areas the SEP law wants to improve (with the exception of teacher salaries, management costs, or infrastructure)[21]. Schools that want to receive the adjusted voucher and the voucher per concentration of priority students cannot charge extra fees to priority students[22].

School accountability. This adjusted voucher is tied to an accountability mechanism in which schools are annually classified by their performance into "autonomous", "emerging" and "in recovery" categories[23]. "Autonomous" schools are those that have shown systemically good results of their students. "Emergent" schools are those that have not shown systematically good results. "In recovery" schools are schools that have shown systematically deficient results.

Schools with different classifications receive different support for their improvement. Schools classified as "autonomous" are required to present a school improvement plan (PME)[24]

[21] The decree DFL 2 from August 20th of 1998 rules the base vouchers of subsidized schools. This decree rules base vouchers can be spent on teacher salaries, salaries of other school personnel, school management costs, costs of infrastructure, infrastructure management costs, acquisition of services and materials for the educational management (pedagogical resources), investment in financial assets as long as the interests are invested on the school, investment on non-financial assets as long as they are required for educational purposes, mortgage payment of the school building, costs with direct relation to the improvement of the school quality, costs with direct relation to the educational plan of the school.

[22] Non-priority students can be charged up to 196 USD; however, schools on average charge 24 USD (Anand, Mizala & Repetto, 2009).

[23] The law indicates no schools can be classified as "in recovery" the first two years the law is implemented.

[24] School improvement plan (PME for its acronym in Spanish for *Plan de Mejoramiento Educativo*) is a document where the school specifies: (i) descriptive of the school, (ii) diagnostic evaluation of the school, (iii) set educational goals, (iv) defined actions to achieve those goals. The goals and actions should consider four areas: curriculum,

which the school communities construct on their own. PME from "autonomous" schools do not need to present a diagnostic evaluation of the school. PME from "autonomous" schools only need to be presented, and not approved by the Ministry. "Emergent" schools need to present such PME[25], but in its construction the principal and school board receive support from the Ministry. Schools "in recovery" have to build the PME with help from the Ministry and a third party technical assistance. The PME of "emergent" and "in recovery" schools has to contain a complete diagnostic evaluation of the school, including an analysis of the latest SIMCE scores, measures of reading speed and reading comprehension in certain grades, an evaluation of organizational management and leadership, and pre-tests on the subjects the school is interested in improving (Elacqua, Mosqueira & Santos, 2009). This PME has to be approved by the Ministry[26].

Schools are offered incentives to increase their quality, both rewards and punishments. Rewards are offered to high performing schools, where the incentive is more autonomy to allocate the extra resources the higher the level of performance of the school. Thus, schools classified as "autonomous" are free to allocate the resources provided by the adjusted voucher. Schools classified as "in recovery" have to allocate 100% of the resources into actions committed by the PME. Schools classified as "emergent" have to allocate 50% of the resources into actions committed in the PME, and the remaining 50% they are free to allocate. Schools

leadership, school climate and resource management. Although the Ministry checks whether schools comply or not with the set goals, the law does not specify what happens if the school does not achieve those goals.

[25] Schools categorized as "emergent" will downgrade to the category of "in recovery" if having adopted the law, they do not present the PME after a year.

[26] Because the PME for "emergent" and "in recovery" schools has to be approved by the Ministry of education, schools report investing a lot of efforts in the elaboration of a proper diagnostic evaluation of the school with help from technical assistance (Barra, 2013).

classified "in recovery" are also subject to sanctions. If such schools do not reach the required standards after four years, the schools get their official recognition revoked, meaning they are no longer eligible for any public funding[27]. Other sanctions include disabling temporarily or perpetually from their functions to the school manager ("*sostenedor*"), 'school principal' or school managers[28].

Schools are classified according to their academic results starting the first year in which they receive the adjusted voucher. The classification scores of the schools depend on their test scores on the national standardized test SIMCE[29] and other quality indicators like retention rates, teachers' evaluations, and integration of the educational community into the educational project of the school.

How schools are classified is specified in Decree 293 and a Technical Report of *Proceso de Clasificacion SEP*. The latter is a report of restricted access provided by MINEDUC through the *Ley de Transparencia*. These documents indicate different sets of rules for the different classifications[30]:

[27] The rule of the 4 year span was postponed by the law 20,529 that creates the Quality Agency of Education (from August 2011). This law indicates the Quality Agency of Education will evaluate schools (starting 2016), and if schools are systematically underperforming for 4 years, then they will get their official recognition revoked. Therefore, no schools should be closed due to their performance before year 2020.

[28] Disabling of "*sostenedores*" and 'school principals' was also postponed until the Quality Agency of Education starts working.

[29] SIMCE is the national standardized assessment conducted by MINEDUC since 1988. The test assesses math, language and social and natural sciences based on the national curriculum. Assessments are performed in 4th, 8th and 10th grade. The scores of SIMCE are standardized on a scale with a mean of 250 points and a standard deviation of 50. This scale was defined in 1999 for 4th grade scores, in 1998 for 10th grade scores and in 2000 for 8th grade scores. This standardization allows analyzing the variation of scores of students over time since 1998, 1999 and 2000 for 10th, 4th and 8th grade respectively.

[30] Appendix A presents the details of the classification rules according to Decree 293 and Technical Report of *Proceso de Clasificacion SEP*.

Schools are classified as "autonomous" if:

A. Rules i to iii should happen simultaneously in at least 2 out of 3 prior SIMCE assessments.
    i. The average school score on SIMCE[31] is above the median of the group of schools with similar SES[32].
    ii. The proportion of students scoring above 250 points is above the median of the group of schools with similar SES.
    iii. The proportion of students scoring above 300 points is above the median of the group of schools with similar SES.
B. The Education Quality Index (ICE index) is above the median of the group of schools with similar SES. The ICE index is constructed by the average SIMCE scores in the last 3 years and other quality indices of the year previous to the classification of the school (see Table 7).

Schools are classified as "in recovery" if:

A. Rules i and ii should happen simultaneously in at least 2 out of 3 prior SIMCE assessments.
    i. The average school score is less than 220 points in SIMCE.

[31] 4th grade SIMCE scores that average math, language and science test scores.

[32] The Quality Agency of Education groups all schools into 4 different groups according to the socioeconomic composition of the students in 4th grade. School's scores are compared to the scores of the schools in their same SES group. The schools' SES level is a measure constructed from four variables: educational level of the mother, educational level of the father, monthly household income and schools' vulnerability index. The three first measures are obtained from a parental questionnaire as part of the SIMCE testing. The fourth measure is a government vulnerability index. This index is constructed by JUNAEB (*Junta Nacional de Auxilio Escolar y Bec*as) to allocate lunch- and health-subsidies to students.

ii. The proportion of students scoring above 250 points in SIMCE is less than 20 percent.

B. The ICE index is in the lowest 10 percentile[33].

Schools that adopted the SEP law, and do not satisfy the above classification rules are classified as "emergent"[34] [35].

The ways schools are classified have advantages and disadvantages. One of the advantages is that the classification looks beyond simple test scores considering also working conditions at the school (SNED improvement score), new initiatives (SNED initiative score), access opportunities (approval and retention rates), integration of teachers and parents in the construction of the school community (SNED integration score), and teacher performance (teacher evaluations). Another advantage is that, the way the evaluation considers test scores is done addressing several drawbacks from ranking schools by simply using raw annual test scores. First, the classification formulae consider several years of performance data. This is a useful way of classifying schools avoiding the confounding influence of mean reversion (Kane & Staiger, 2002; Chay, McEwan & Urquiola, 2005). Second, schools are not classified based on absolute test scores but on reference to homogeneous groups in terms of SES. In Chile there is a strong correlation between schools'

[33] Schools are ranked according to the ICE score. Those schools that are in the 10 percent lowest ICE do not satisfy this criterion.

[34] The classifications of some schools do not follow the formulae explained above. Some of these are: (i) schools with less than 20 students taking the SIMCE test in 4th grade; (ii) multiple-grade schools, geographically isolated schools, and schools with three teachers or less; (iii) schools with less than two standardized national evaluations in previous years (e.g., all new schools); (iv) schools that do not have 4th grade.

[35] Schools that do not have 4th grade are classified as "emergent" regardless of their academic results. This situation was not explained in the law on 2009. Such law mandates the gradual introduction of different grades. It starts in 2008 with PK to 4th grade, and then each following year it introduces a new grade. Supposedly this meant that as soon as 8th and 11th grade were participating in the law, the SIMCE scores of those grades will become high stakes (that was explicitly mentioned in one version of the law for the case of 8th grade). That did not happen. Only 4th grade test scores are high stakes. 8th and 11th grade test scores will become high stakes once the Quality Agency of Education starts classifying schools in 2016.

test scores and schools' SES (Mizala, Romaguera & Urquiola, 2007). If the classification was done using raw scores, then the classification would simply resemble the SES level of the schools. Classifying schools in reference to similar schools avoids a full correspondence between classification and the SES of the school.

One disadvantage is that the formulae are quite complex in terms of how all the indicators are related, and it may not be easy for schools, teachers and parents to understand why a school was classified the way it was. Furthermore, the formulae are quite recursive using the same scores dressed up with different clothes. For example, SIMCE scores are used as an average, separately it is considered the proportion of students scoring above 250 and 300 on SIMCE scores, and SIMCE scores are also the main component of the ICE index. Another disadvantage of the performance measure available is that SIMCE scores do not allow tracking students over time, and, therefore, it is not possible to construct proper value-added measures.

The information about the classifications of schools is given by MINEDUC to the school owner ("*sostenedor*") and they inform the 'school principals' and managers. They also have to inform parents about the school classification. MINEDUC also publishes the school classifications on its website.

Implementation rules. Adoption of the SEP law is voluntary. To have access to the extra resources the school signs an agreement of equality of opportunities and academic excellence (CIOEE) where they commit not to select students by prior academic achievement (until 6th grade), improve school results, retain students, not charge priority students any top-up fee, and

design and implement a PME[36]. This agreement lasts for four year. It can be renewed if schools want to.

The law mandates implementation from the first year from PK to 4th grade, and then every year after a new grade is introduced. Thus, in 2008 Pk-4th grade will be embraced by the law, in 2009 it will add 5th grade, the next year 6th grade, and so forth. In 2016 the implementation of the law is complete as it incorporates 12th grade.

3.3. Implementation

School adoption to the law was gradual. The first year, in which the SEP law was in place, 96.75% of public schools that offer elementary education adopted it, whereas only 42.30% of private voucher schools did. By 2010, 99% of public schools had adopted it, and 54.6% of private voucher schools (see Table 8). The different rates in which schools adopted the law is a clear response to fact that public schools have a much higher proportion of priority students than private voucher schools[37].

[36] This agreement also requires the schools to [Art. 7]: (i) Present an annual report about the use of resources acquired by the adjusted voucher –all resources received must be reported; (ii) Accredit the school board, teachers' board, and parental board; (iii) Credit the existence of teacher's time to pedagogical advisement and planning in the school and teacher's academic time to non-lecture activities such as class planning; (iv) Present an Improvement plan of the school (PME), elaborated with the school community (this plan should consider areas of curriculum, leadership, school climate, and management of school's resources [Art 8]); (v) Define and meet goals of academic achievement of their students, especially priority students. These goals should be defined using the national standardized testing system (SIMCE); (vi) Indicate in the agreement the total amount of public resources received. This data should be presented annually; (vii) Make public for parents and students this agreement, especially what refers to the academic goals of the school; (viii) Safeguard that all teachers present the academic annual plan to the school's principal during the first 15 days of school; (ix) Include in the academic planning cultural, artistic and sport activities.

[37] In 2008, 35.58% percent of students enrolled in public schools where priority students (eligible for the adjusted voucher), in contrast to 17.24% in private voucher schools. In 2010, 58.91% of public school students were eligible; whereas in private voucher schools it was only the 36.06% (see Table 9).

The classification schools received was gradual. During the first four years schools were only classified as “autonomous” and “emergent”. Supposedly by 2010 schools should have also been classified as “in recovery”; however, that did not happen. Schools started being classified as “in recovery” only by 2012. According to staff from MINEDUC, there is no memorandum or record of why this delay happened. By 2012, 7,459 schools received a classification. However, most of them received a classification of “emergent” because they either had less than two SIMCE measures or because the schools have had 20 students or less taking the tests (average of the three tests in the last three measures). Thus, only 39% was classified using the classification formulas; percentage that has slightly decreased with the years (see Table 10).

Table 11 shows the distribution of schools in the different categories for the first four rounds of complete school classification, for all schools classified using the formulas. The greatest variation across years is in the lowest classification, where there is a steady decrease in the number of schools classified as “in recovery”. The variation of schools classified between “autonomous” and “emergent” is not quite large on average. This is something to be expected considering the classification rules are relative to the median test scores of the SES groups. In Table 12, I disaggregate such broad number in the number of schools that change from one category to another in the following years. The way to interpret this table is as follow. For example, from all schools that were classified as “autonomous” on year 2012, 809 were classified as “autonomous” on 2013, and 243 were classified as “emergent”. From all schools classified as “autonomous” on 2013, 848 were classified as “autonomous” on 2014 and 209 were

classified as “emergent”. And so forth. Overall, the table shows there is variation across years in the number of schools that fall into each classification[38].

Information about how schools are classified is not easily available. Although the criteria used to classified schools is reported in the Decree 293, how exactly these criteria are calculated and which are the threshold used are not fully reported there. Such information is contained in the Technical Report of *Proceso de Clasificacion SEP* which can only be obtained by filing a document request under the *Ley de Transparencia*. This suggests that schools are not likely to game the system, unless they had asked for and reviewed this report.

Information about the classifications the schools get is published in on MINEDUC’s website (www.mime.mineduc.cl). In this website parents can find several indicators of all schools. Information about school classification in the SEP scheme is hidden under the ‘Cost Indicators’ section. However, what each classification means is not explained, and it does not say the classification refers to school quality. From the context, each classification could be interpreted as some indicator of ‘costs management’ or alike. There is also no data regarding how schools inform parents about the classification the school gets.

The use of resources has been a matter of public discussion. Reports from the government entity dedicated to the monitoring and control of public administration expenses have uncovered thatthis information of all the funds received by schools in 2011, 37% were not appropriately accounted for (Contraloria General de la Republica, 2014). 29.4% of incoming resources were

[38] Note that in the table the vertical sum of one category does not necessarily match the horizontal sum of the category the year after. This happens because not all schools participate in the SEP law every year (some schools did not renew their CIOEE after 4 years of participation), and because some schools were classified according to the formulas some years but some other were not.

used for things not approved by the law, and 7.4% of incoming resources were missing. It would not be a surprise if the situation was somehow similar for other years.

3.4. Empirical evidence of the SEP law

If the SEP law had a positive impact on student achievement, we would expect SIMCE scores to increase starting in 2008. Figure 2 shows standardized math and reading achievement scores in Chile from year 2006 to 2013. The figure shows the academic results of 4th grade students in math and language have a steep growth in the period 2008 to 2012, after years of stagnation. However, we cannot causally attribute this increase in scores to the introduction of the SEP law. There could be other time-varying factors influencing the results.

In the last couple of years several studies have assessed the causal impact of the introduction of the SEP law on educational quality (see Table 13). Overall the findings indicate that the SEP law has mildly increased schools' high-stakes test scores (MINEDUC, 2012; Villarroel, 2012; Correa et al., 2013; Mizala & Torche, 2013; Neilson, 2013; Navarro-Palau, 2015), and particularly those of vulnerable students (Neilson, 2013; Navarro-Palau, 2015). However, no studies have assessed the impact of the SEP law on other academic outcomes (e.g. low-stakes tests, retention rates), nor it is clear whether the effect is due to the extra resources schools are receiving or due to the school accountability mechanisms set in place (i.e., more government control over school performance).

In order to assess whether the SEP law increases school quality, recent studies take advantage of the fact that the adoption of the law is voluntary and private voucher schools have adopted it gradually (MINEDUC, 2012; Correa et al, 2013; Villarroel, 2012; Mizala & Torche, 2013). The comparisons of academic outcomes between schools that adopted the law and those

that did not, offer an interesting counterfactual. Unfortunately this type of analysis is only possible to do with private voucher schools and not with public schools because almost all of the latter adopted the law during the first year of the implementation of the law; therefore, it was not possible to compare those public schools that adopted the law versus those that did not. In the case of private voucher schools, however, only 45% of the schools adopted the law by 2012. This meant 55% of private voucher schools that did not adopt the law, and could serve as a control group. All the studies that used this criterion to gain identification (MINEDUC, 2012; Correa et al., 2013; Villarroel, 2012; Mizala & Torche, 2013) used panel data from 2006 to 2011 aggregated at a school level; the outcomes assessed were math and language 4th grade test scores.

Villarroel (2012) studied the impact of the SEP law on academic outcomes of 4th graders. He compared private voucher schools that had and had not adopted the law in 2010. To control for selection bias of some voucher schools adopting the law, the author used propensity score matching to control for the probability of selection into the SEP law. He then used a difference-in-differences method with two time periods to estimate the impact of the policy (2007 and 2010). He found a positive impact on language and math scores, with effect sizes ranging between 0.11 and 0.18 in math and 0.07 and 0.11 in language. The impact was larger for schools with a higher proportion of priority students and for schools that had implemented improvement plans for a longer time.

In 2012, the Center of Studies of MINEDUC and the Department of Studies of the Ministry of Finance released an impact evaluation after four years of the beginning of the implementation of the SEP Law. The study used a difference-in-differences approach to estimate the effect of adoption of the law among private voucher schools. Different from Villarroel (2012), this paper used more than two time periods; the study used a panel of data from 2006 to 2011. The authors

found the SEP law had a positive and significant effect on academic achievement of fourth graders of private voucher schools, with effect sizes ranging from 0.08 to 0.11 in math and 0.05 to 0.07 in language[39]. Each extra year participating in the SEP law increased the achievement by 0.03 in math and 0.02 in language[40] [41].

The study by Mizala and Torche (2013) analyzed the effect on academic achievement of schools' participation on the SEP Law, the impact of participation over time, and heterogeneous effects on schools according to schools' socioeconomic background. The authors used the sample of private voucher schools because there is variation in the timing when these schools adopted the law. The authors used a fixed effects model where they use a fixed effect per school and per year. This strategy allowed them to identify the within school difference due to changes in the status of treatment. They also controlled for changes in the socioeconomic composition of schools. The authors found that private voucher schools increased their test scores once they adopted the SEP Law. The effect sizes in math ranged from 0.08 to 0.1 and from 0.07 to 0.08 in language[42]. Each extra year of participation in the SEP law impacted positively on the outcomes. This suggests schools require time to adjust their resources, pedagogical strategies and teaching plans. The authors also analyzed the effects by the average socioeconomic level of the families of students enrolled in schools. The authors found schools from lower quintiles had a higher increase of their test scores than schools from higher quintiles. Schools from higher quintiles had

[39] Effect sizes calculated dividing the reported coefficients by the standard deviation of SIMCE (50).

[40] This study adds years in the models as a continuous variable and not as fixed effects. This assumes that the effect of each year is the same on students' outcomes. This assumption is relaxed by the study of Mizala and Torche (2013).

[41] There is a more extended version of this study by Correa, Inostroza, Parro, Reyes and Ugarte (2013).

[42] Effect size calculated dividing the reported coefficients by the standard deviation of SIMCE (50).

an effect that was negligible. This suggests the effects are aligned with the objectives of the policy as it expects to impact the least advantage population of students.

The advantages of the fixed effects models for school and time used by Mizala and Torche (2013) is that such strategy allows them to control for unobservable variables that vary across schools but not over time (school fixed effect) and vary over time but not across schools (time fixed effect). There are two limitations of these strategies that may bias the findings. One of the limitations refers to the omission of those variables that change within units over time. Schools could change their activities to accommodate the extra resources and to accommodate priority students. Also students may sort between schools. Such a situation could happen if the compositions of schools change as they adopted the law (as they move from being in the control to the treatment group). Changes in observable characteristics of the composition of schools are controlled for as characteristics of the student group are added into the models, e.g., average SES of the students in the school, average parental education, etc. Changes in unobservable characteristics, however, are not addressed. Some unobservable characteristics that modify the composition of the schools could appear if parental choice is determined by the fact that schools adopted the SEP law. It is possible that parents that have certain special motivations or interests choose schools that adopted the SEP law. If this is the case, then it is possible the findings are biased upwards. The second limitation is that the fixed effects model only estimates the impact of the law on schools that exist before and after the law. Therefore, variations offered by schools that leave or join the market are not accounted for. The law could have impacted the number of schools by (i) attracting new schools to join the market, as now there is more money to run the schools, or (ii) somehow pushing away schools (although it is hard to think about this possibility as schools that do not want to join the law can do so). The direction of the bias could be both

ways. It could be upward biased if new schools start with low academic achievement as they get everything up and running. Or it could be downward biased if the law somehow pushed away from the market low performing schools.

One of the latest studies assessing the impact of the SEP law takes a slightly different approach and assessed the impact of the law on students of both public schools and private voucher schools. Neilson (2013) used a difference-in-differences approach comparing the impact of the law on poor and non-poor students across the years (intent-to-treat). He used panel data from 2004 to 2011 at a student level; test scores used are those of math and language for 4th grade students. He found that the targeted vouchers raised poor student test scores by 0.2 standard deviations. He also found that the targeted vouchers close the gap between the 40% poorest population of students and the rest of the students by one third. The author simulated a demand and supply model to decompose the effects of the policy, specifically to identify whether the increase in test scores was driven by an increase in the quality of the schools or by sorting of students. He found the supply reaction accounts for two thirds of the effect, whereas the demand reaction accounts for one third of the effect. The author attributes the increase of school quality as a reaction of schools to increased competition for vulnerable students.

The latest study of the introduction of the SEP law on high-stakes outcomes is from Navarro-Palau (2015). She uses a novel regression discontinuity with difference-in-differences design to estimate the impact of increased school choice due to the introduction of the law on 4th grade SIMCE scores. Overall she finds there is no evidence the increased school choice improved test scores. However, when a maternal education moderator is considered, she finds children of less educated mothers (those who tended to stay in public schools) had an increase in average test scores of 0.08 standard deviations. There is no evidence of impact on children of more educated

mothers (those who tended to switch to private voucher schools due to the introduction of the law).

None of the studies mentioned above considers explicitly that the SEP law has an accountability component, a component that could have triggered the impacts on high-stakes outcomes. Therefore, there is an important question that still remains unanswered, whether the impact of the SEP law would have been the same if the accountability component had not been in place.

3.5. Conclusion

The Chilean educational system has a mix of market policies, where there is a marketplace of schools with vouchers per enrolled student and vouchers per vulnerable student that are conditional on schools' performance. The latter vouchers were introduced by the SEP law as a solution for the unintended consequences of having a system of flat vouchers per enrolled student. Whether they have increased the quality of schooling received by priority students is the topic of this study.

The studies that assess whether the SEP law has impacted on school quality conclude it has had positive effects on schools' high-stakes test scores. On average, schools participating in the SEP law have been estimated to increase mathematics test scores by 0.1 standard deviations and language test results by 0.08 standard deviation (Villarroel, 2012; MINEDUC, 2012; Correa et al., 2013; Mizala & Torche, 2013). Student level data indicates that the SEP law raised test scores of the poorest 40% of students on average by 0.2 standard deviations (Neilson, 2013).

All the current studies analyze the effect of the SEP law as a whole, as a black box. None of them tries to separate the effect of the adjusted vouchers (i.e., extra resources) from the effect of

the accountability component of the law (i.e., classification of schools and resource allocation autonomy conditional on performance, and more information that can affect parental school choice). Although Neilson (2013) looks at the supply and demand responses to the policy, he suggests the quality of schools may have been altered only by having schools compete for more resources. It will be interesting to know how much of the increase in quality is due to competition and how much is due to the accountability component of the law. Disentangling the effects of the law is crucial for informing educational policy. Would the impact of the law have been the same if only extra resources for vulnerable students were provided[43]? Or if the accountability component was the only aspect of the law set in place? Had there been any effects if only more information about school performance was provided[44]? Or were the explicit consequences of resource management what drove schools to improve their academic results? Furthermore, how the introduction of the law affects the school processes may indicate how the law is affecting the academic outcomes of the students on high-stakes tests. These evaluations are important as they inform policymakers not only about the impact of the SEP law, but about the mechanism in place.

[43] Evidence of the impact of differentiated vouchers for disadvantaged students on the Netherlands has shown negative and significant effects on students test scores of the whole student population (Leuven, Lindahl, Oosterbeek & Webbink, 2007). However, other studies addressing the causal effect of extra resources on students' outcomes have shown positive impacts on test scores (Lafortune, Rothstein & Schanzenbach, 2016) and on long term outcomes (Jackson, Johnson & Persico, 2015), specially for low-income districts.

[44] As it is found in other studies of the impact of school accountability, that address the impact of providing extra information (without consequences enacted by the government) (Andrabi, Das, & Khwaja, 2014; Witte, Wolf, Cowen, Carlson & Fleming, 2014). For example, Witte and colleagues (2014) study the impact of an accountability system on the Milwaukee voucher system. This accountability system only provides extra information for parents and schools. In fact, this is the first time parents are informed about individual schools outcomes. The state's central educational agency had statutorily prohibited from sanctioning or rewarding schools based on their results. However, new information could trigger consequences in the choices parents make. The authors found a positive and significant impact one year after the implementation of the accountability system.

All the impact evaluations of the SEP law have focused on the academic measures of students in the high-stakes test SIMCE, particularly in 4th grade math and language scores. To my knowledge, there are no evaluations of the impact of the SEP law on low-stakes tests nor on long-run outcomes. Low-stakes test and long-run outcomes may reflect other aspects of the desired outcomes of the law, which is to improve school quality. Using non-incentivized performance measures to evaluate the impact of an accountability policy is relevant as impact on high-stakes tests can be simply a reflection of score inflation (Koretz, 2005; Jacob, 2007; National Research Council, 2011; Lee, 2008; Chiang, 2009).

Furthermore, it is important to assess whether the incentives result in a sufficient increase in the desired output to justify the costs of running the incentives system (National Research Council, 2011).

To my knowledge there are no evaluations of the impact on academic outcomes of accountability pressure, for high-stakes, low-stakes or long-run outcomes. As far as I am aware, there is only one evaluation of the impact of the law on educational processes from Elacqua and colleagues (2015), and it specifically assessed the impact of accountability pressure on low-performing schools. They found low-performing schools decrease teacher training, decrease crossed-observation of classes among colleagues, increase after school tutoring, and hire teachers with increased teacher's test scores from their undergraduate exit exams (Elacqua, Martinez, Santos & Urbina, 2015).

In sum, the evidence suggests the SEP law has a positive but modest impact on high-stakes tests. No studies have assessed the impact of the law on low-stakes tests. Nor have any studies

attempted to disentangle the effect of the extra resources provided by the law and the effect of the school accountability mechanism.

In the next chapter I present the methodology and data used to assess the impact of accountability pressure on academic outcomes and potential mechanisms of the SEP law.

Chapter 4 - Methodology

4.1. Research questions and hypotheses

The purpose of this dissertation is to assess the impact of school accountability in Chile. I do not attempt to estimate the systemic impact of the accountability system, i.e., the introduction of school accountability in Chile. Rather, I wish to try to separate out estimates of the impact of accountability pressure, i.e., the pressure schools face as they receive different classifications.

I explore three questions about the ways in which school accountability affects students' educational outcomes:

(1) What is the impact of accountability pressure on high-stakes outcomes? The basic prediction of the incentive theory is that schools threatened by a punishment or by losing a reward will improve their performance in the subjects and grades in incentivized outcomes.

(2) What is the impact of accountability pressure on low-stakes outcomes? The theory of incentives predicts that multitasking agencies will focus their attention on the rewarded outcomes (high-stakes outcomes) and not in other outcomes (low-stakes outcomes) if they are not complementary to the rewarded outcomes (Holmstrom & Milgrom, 1991). Given that consequences are attached to the performance of the school in some grades, one could expect that schools shift their resources and efforts towards those grades included in the school classification rules (high-stake outcomes in high-stakes grades), instead of the outcomes in other grades (low-stakes grades).

(3) If school accountability increases incentivized or non-incentivized outcomes, what mechanisms drive those improvements? Understanding the potential mechanisms that drive the policy impact (or the null impact) could have extensive policy value (Ludwig, Kling &

Mullainathan, 2011). It can help us understand what aspects of the policy (or the context in which it is implemented) may moderate its impact.

As mentioned earlier, the mechanisms through which school accountability seems to operate to increase performance measures depend on (i) the behavioral response of the school, (ii) the student body composition, and (iii) the teacher body composition. In terms of school's behavior, the accountability pressure may generate both desirable and undesirable school behaviors. If the adopted behavior increases the performance measure at the expense of the educational goal, then the behavior is undesired. However, if the behavior improves the performance measure as well as the educational goal, then the behavior is desired. In terms of student mobility between schools, information about the quality of schools may affect the way parents choose schools for their children, or it may affect the way schools select their students, or both. Schools classified as low performing may be less attractive for parents, and/or these schools may intentionally select more prepared students to improve their results. Furthermore, schools may have an incentive to increase their prices, or to increase the number of students from low SES background as the school will get more money but risking lowering school results. In terms of teacher mobility, schools classified as low performing may find it challenging to retain and hire teachers of good quality.

4.2. Empirical strategy

The empirical method I use follows the work of those who study the impact of accountability pressure (e.g., Figlio & Rouse, 2006; Rockoff & Turner, 2008; Chiang, 2009; Deming, et al., 2013; Mizala & Urquiola, 2013; Clotfelter et al., 2004; Hart & Figlio, 2015; Elacqua, et al., 2015), and those who have used regression discontinuities when there are multiple rating scores

(e.g., Reardon & Robinson, 2012; Papay, Willet & Murnane, 2011; Robinson, 2008; Cohodes & Goodman, 2014; Papay, Murnane & Willett, 2014; Elacqua et al., 2015).

The causal effect of the classification of the school on the school outcomes could be done comparing the outcomes of schools that received a classification of "autonomous" with those that received the classification of "emergent" only if schools assigned to each category were done randomly. But this is not the case. If we compare the outcomes of "autonomous" schools and "emergent" schools we would confound the impact of the school categorization with the fact that "autonomous" schools perform better than "emergent" schools initially. In order to account for such omitted variable bias I use a regression discontinuity design (RD hereafter). An RD allows me to compare the outcomes of schools just above and below the school classification threshold, as schools should be similar except for receiving one classification or the other.

The RD design exploits natural discontinuities in the rules used to assign individuals to the treatment to compare the outcomes of the individuals assigned and not assigned to treatment. Using the potential outcomes framework (Holland, 1986) this can be phrased as follows. Let there be a rule that indicates that above certain score (c) of a rating variable (X) individuals will receive a treatment $(D = 0)$, and if they score below such score individuals will receive treatment $(D = 1)$. Each individual has two potential outcomes: one that results from the individual being assigned to treatment (0), $(Y_i(0))$, and a second one from being assigned to treatment (1), $(Y_i(1))$. What we want to know is $E[Y(1)|X = c] - E[Y(0)|X = c]$. However, we do not observe $Y(0)$ on $X = c$. Under the assumption that $E[Y(0)|X = x]$ and $E[Y(1)|X = x]$ are continuous in $(x = c)$, we can consider the average treatment effect of treatment (1) relative to (0) being

$$E[Y(1) - Y(0)|X = c] = \lim_{x \uparrow c} E[Y|X = x] - \lim_{x \downarrow c} E[Y|X = x] \quad (1)$$

which is interpreted as the average causal effect of the treatment at the discontinuity point (Imbens & Lemieux, 2008).

The estimate of this difference can be obtained using a parametric regression model[45],

$$Y_i = \alpha + f(x_i) + \rho\, D_i + Z_i\varphi + \varepsilon_i\,, \quad (2)$$

Where x_i is centered at the cutoff score, f is a continuous function at $x = 0$, and D is a dummy variable indicating whether $x > c$ or not. ρ is our coefficient of interest. Z is a vector of covariates. The later should not radically change the estimates (Imbens & Lemieux, 2008). The main purpose of including covariates is to increase precision.

In this study, the treatment status is not determined only by one rating score, but several rating scores. For example, schools are classified according to their SIMCE scores, proportion of students achieving above 250 points and 300 points in SIMCE, etc. For this reason I use a variation of the RD for multiple rating scores.

Furthermore, in this study, there are not only multiple rating scores, but there are two separate sets of administrative rules that each determine two treatment statuses. As mentioned earlier, in the SEP law, schools with scores above or below multiple thresholds are classified into three different categories: "autonomous", "emergent" or "in recovery". The classification rules to define which schools are "autonomous" versus "emergent" differ from the rules to determine which schools are "emergent" versus "in recovery". The rules to determine if a school is

[45] Another possibility is to use a non-parametric approach by taking the difference between the average of individuals' outcomes right above and below the cutoff. A third possibility is a semi-parametric approach by estimating separate linear functions just above and below the cutoff, and taking the difference in predicted values at the cutoff.

"autonomous" are relative to schools that have a student population of similar socioeconomic background. For example, a school with a student population of SES level A will be compared to the median scores obtained by schools of SES level A. However, the rules to determine if a school is "in recovery" are relative to absolute thresholds. For example, a school will be compared to a threshold of 220 points in a test score, regardless of the SES level of the school. "Emergent" school will be all those that are not "autonomous" nor "in recovery".

Given there are two separate sets of rules to classify schools, I analyze the effect of the categorization of schools separately. First I analyze the impact of receiving a category of "autonomous" versus the other categories. Then I analyze the impact of receiving the classification of "in recovery" versus the other categories.

In this study I use a **fuzzy regression discontinuity**, which combines regression discontinuity and instrumental variables (IV) approaches. The IV approach is needed because achieving all cutoff scores does not perfectly predict the school classification. This happens for two reasons. First, because it is not each cutoff score that matters, but some of them need to be achieved simultaneously in at least two years, and others need to be achieved in just one year. Second, the way the rules are defined is not straightforward, but there are several ways of interpreting the rules[46].

[46] To illustrate this point, in Appendix C, in Table C.1, I show the level of coincidence between the classifications as MINEDUC publishes them and the classifications as I calculate them according to my interpretation of the rules specified in the law, decree and technical documentation (Appendix A shows the details of such interpretation of rules). Overall, the level of coincidence varies between 83 and 100 percent in the different classification categories and years. Some of the issues of the rules that are subject to interpretation are: a) how is the median of the SIMCE, proportion of students scoring above 250 and 300 points, and ICE scores of the school by SES calculated, either using all schools or only classifiable schools; b) how are complementary indicators standardized, either using all schools or only classifiable schools; c) how many students must take the exam in order for the results of the school to count as classifiable, 20 or more, or more than 20; d) whether the SIMCE scores and proportion of students scoring highly must be attained in at least two years, either all the rules at least two years or any rule at least two

Estimation of the effect of receiving the classification of "autonomous"

I estimate the model using as an IV the score that determines more strongly the classification of schools. To find which one is the binding score on the classification of "autonomous" schools on 2012 I run several models where the outcome is whether the school is "autonomous" or not, and using each rule of the classification as the determinant.

In Table 14, I present such models. Independent variables called "above threshold" are dummy variables where 1 indicates whether the school scores are above the median of the SES group in the specific rule mentioned in the top of the columns. The variable called "running variable" is the rule mentioned in the heading of the column centered at the SES group median. The models are estimated using linear probability models and they all include covariates such as the type of school, whether the school is urban, socioeconomic level of the school and 4th grade enrollment rates.

The models indicate that the score that determines most strongly whether the school receives the classification of "autonomous" is the ICE score the school got the year prior to the classification. The strength of the relationship is measured by the coefficient of the dummy variable that indicates whether the school scored above the threshold, and by the F-test of the model. The coefficient of the dummy variable of the model of column (10) indicates that being above the SES group's threshold explains about 53 percent of the classification the schools get. This coefficient of the ICE index is up to two times larger than the coefficients of the other scores that determine the classification the school gets. The F-test of the model in column (10)

years. Despite trying the classification of the schools with all these combinations of rules, I never achieved total match between the classification I can predict and the one MINEDUC publishes.

indicates this model fits best the population of schools than the others that use other scores as determinants of the classification of schools.

To estimate the causal effect of the school classification of "autonomous", I use as a first stage a linear probability model to predict the classification schools receive (Q) whether it is "autonomous" ($Q = 1$) or not ($Q = 0$).

$$Q_{st} = \alpha_0 + \alpha_1 ABOVE_{s(t-1)} + \alpha_2 ICE_{s(t-1)} + \alpha_3 ABOVE_{s(t-1)} * ICE_{s(t-1)} + X\alpha_4 + \mu_{st} \quad (5)$$

$ABOVE_{s(t-1)}$ is the instrumental variable, where $ABOVE = 1$ if the school has a ICE index above the median of the SES group ($ICE_{s(t-1)} \geq 0$), and $ABOVE = 0$ otherwise ($ICE_{s(t-1)} < 0$). X is a vector of school characteristics (i.e., type of school, SES of the school, rural/urban, enrollment rates).

Then, I use the predicted value of Q (i.e.,$\hat{Q}$) estimated from the equation 5 to predict the school outcomes on the following years to the classification ($t + k$).

$$Y_{s(t+k)} = \beta_0 + \beta_1 \widehat{Q_{st}} + \beta_2 ICE_{s(t-1)} + \beta_3 ABOVE_{s(t-1)} * ICE_{s(t-1)} \quad + X\beta_4 + \varepsilon_{s(t+k)} \quad (6)$$

Thus, β_1 is the unbiased estimate of the local average treatment effect of being classified as "autonomous"[47] [48]. This local effect is only estimated for schools that comply with the rule of the

[47] Given the binary nature of the treatment, a probit in the first stage could have been appropriate. However, plugging the predicted value of such first stage onto the second stage would be estimating what Angrist and Pischke (2009) called the "forbidden regression". This regression refers to the direct application of 2SLS to a nonlinear model. The problem is that only OLS estimation of the endogenous variable is guaranteed to produce first-stage residuals that are uncorrelated with fitted values and covariates. In cases like this, Wooldridge (2002) suggests adding an extra step where we run an OLS of Q on the predicted value of the probit ($\hat{Q}$) and all the other covariates. The estimated values of this OLS ($\widehat{\hat{Q}}$) are plugged into the second stage (equation 6). However, the disadvantage of this proceeding is that it uses the predicted classification status rather than the actual autonomous status as the endogenous second stage predictor.

[48] Both continuous and dichotomous outcomes are calculated using OLS.

score that is binding (the instrumental variable restriction to generalizations). This estimand is valid only for schools that are close to the passing threshold (the regression discontinuity restriction to generalizations).

Estimation of the effect of receiving the classification of "in recovery"

To estimate the impact of schools being classified as "in recovery" I follow the same rationale used to identify the impact of the "autonomous" classification. I identify the binding score for the "in recovery" classification, and then use such binding score as IV to estimate the impact of the classification on the outcomes of the schools. There are two small differences with the model described for the "autonomous" classification. First, because there are too few schools classified as "in recovery" (Table 11), I use two rounds of classifications pooled together, the rounds of year 2012 and 2013. I add a fixed effect per year to avoid any confounding impact of time varying unobservable variables. Second, instead of defining the IV as "above", I use a "below" IV because the classification of "in recovery" is for those schools that score below the different cutoffs.

In Table 15 I present the resulting models in the search for the binding score of the "in recovery" classification. Independent variables called "below threshold" are dummy variables where 1 indicates whether the school scores are below the cutoff in the specific rule mentioned in the top of the columns, zero otherwise. The variable called "running variable" is the rule mentioned in the top of the column, centered at the cutoff score. The models are estimated using linear probability models, they all include covariates (type of school, whether the school is urban, socioeconomic level of the school and 4th grade enrollment rates), and a fixed effect per year.

The estimated coefficients indicate that the score that more strongly determines whether a school is classified as "in recovery" is the ICE score the school received the year prior to the classification (Table 15, column 7). Again the strength of the relationship is measured by the coefficient of "below threshold" and by the F-statistic of the model, both considerably larger than the coefficient and the F-statistic of the models in the other columns. The coefficient of the "below threshold" dummy indicates that being below the cut score explains about 97% of the classification the school gets. This is twice as much as the coefficients of the other models.

To estimate the causal effect of the school classification of "in recovery", I use as a first stage a linear probability model to predict the classification schools receive, whether it is "in recovery" ($R = 1$) or not ($R = 0$).

$$R_{st} = \gamma_0 + \gamma_1 BELOW_{s(t-1)} + \gamma_2 ICE_{s(t-1)} + \gamma_3 BELOW_{s(t-1)} * ICE_{s(t-1)} + X\gamma_4 + \delta_t + \mu_{st} \quad (7)$$

$BELOW_{s(t-1)}$ is the instrumental variable, where $BELOW = 1$ if the school has an ICE index below the 10th percentile cut score ($ICE_{s(t-1)} < 0$), and $BELOW = 0$ otherwise ($ICE_{s(t-1)} \geq 0$). X is a vector of school characteristics (i.e., type of school, SES of the school, rural/urban, enrollment rates). δ_t is a year fixed effect.

Then, I use the predicted value of R (i.e., $\hat{R}$) to estimate the impact on the outcomes of the following years to the classification ($t + k$).

$$O_{s(t+k)} = \vartheta_0 + \vartheta_1 \widehat{Q_{st}} + \vartheta_2 ICE_{s(t-1)} + \vartheta_3 BELOW_{s(t-1)} * ICE_{s(t-1)} \quad + X\vartheta_4 + \delta_t + \varepsilon_{s(t+k)} \quad (8)$$

Here, $O_{s(t+k)}$ is an outcome measure of the school s on year ($t + k$). The identifying assumption of this equation is that any relative increase/decrease in school outcomes of those

schools that where classified as "in recovery" induced by missing the cut score of the ICE score is attributable to the accountability classification itself. If the identifying assumption is correct, ϑ_1 is the local average treatment effect of being classified as "in recovery". The effect is valid only for schools close to the threshold.

4.3. Data

Sources of data

To conduct the analyses I use school level data from three different sources: administrative data and school classification scores from the Ministry of Education (MINEDUC), standardized test scores from the Quality Agency of Education (in Spanish *Agencia de Calidad de la Educacion)*[49], and administrative data from the Chilean Internal Revenue Service (in Spanish *Servicio de Impuestos Internos*).

Administrative data from all schools is publicly available on the MINEDUC website. The administrative dataset of MINEDUC provides data on the characteristics of the school, the number of enrolled students per grade, number of students with special needs, number of priority students (i.e., students eligible for the adjusted voucher), number of beneficiary students (i.e., students eligible for the adjusted voucher who are enrolled in a grade and school that has adopted the SEP law), students' fees (reported by schools). Data from the internal revenue service provides information about which schools are for-profit and which ones are not[50].

[49] This study used data from *Agencia de Calidad de la Educacion*. The author is thankful to *Agencia de Calidad de la Educacion* for allowing access to the data. All the results of the study are responsibility of the author and not the *Agencia de Calidad de la Educacion*.

[50] Schools are said to be "potentially for profit", which does not necessarily mean they make profits. There is no data available about how much profit they make, if any. This data is not available for year 2012. It is available for 2013.

Data on the SEP classification of schools comes from the Department of National Coordination of School Subsidies from the Ministry of Education[51]. The data about the SEP classification of schools contains data on schools that have joined the SEP law from 2012 to 2015. Data is available for each one of the rating scores the school gets and the final classification of the school. The data also allows identification of which schools were classified due to their quality measures and which ones were classified due to administrative criteria (e.g., not presenting the PME after a year of joining the SEP law).

Data on the educational outcomes of schools and the socioeconomic background for their students comes from the Quality Agency of Education. The data on educational outcomes of the schools are the 4th and 8th grade test scores of SIMCE (see footnote 28 for more details)[52]. Accompanying the scores is the assessment of parents, teachers and students from a survey. These surveys provide information about the socioeconomic characteristics of the students among other measures.

Sample

The universe of schools consists of SEP schools that provide elementary education[53]. There is data for 7,459 schools on 2012 and 7,967 schools on 2013. From all those schools, around 62 percent are excluded from the sample because they either have less than 20 students or have less than two measures of academic outcomes, and therefore are not classified using the classification

[51] I am grateful for the help provided by Ignacio Monge from the Department of National Coordination of School Subsidies from the Ministry of Education to have access to the data.

[52] Recently in 2012 the Ministry started evaluating 2nd and 6th grade students as well. I have omitted those evaluations from my research because these assessments are still in their early stages. The calendar of evaluations is in Appendix B.

[53] Schools that do not provide 4th grade courses are not subject to classification rules, but are classified as "emergent" regardless of the academic results of the school. Elementary education in Chile considers grades 1st to 8th.

formula (see Table 10). I also exclude 24 schools because their identification codes and verification codes for identification are inconsistent throughout the years, and therefore it is impossible to match their outcome data. I also exclude five schools that in 2011 were classified as "emergent" and in 2012 comply with all the rules to be classified as "emergent" but where classified as "in recovery". I suspect these schools were classified as "in recovery" not due to the use of the classification formula, but because they did not present the PME (this could be the cases for which I have missing data –as seen in Table 10). The potential sample is composed of 2,901 schools that in 2012 are classified using the classification formula (39% of all schools; see Table 10), and 2,867 in 2013.

I use two analytical samples: one for the analysis of the impact of the "autonomous" classification, and one for the analysis of the impact of the "in recovery" classification.

For the analysis of the impact of the "autonomous" classification I use cross-sectional data for the 2012 round of classification. My analytical sample uses a bandwidth of 0.706, reducing the sample to 2,212 schools. Descriptive statistics of the schools in the sample are presented in Table 16. The table presents the average characteristics for schools above and below the cutoff. The distribution of schools above and below the cutoff differs across urban and rural schools and across municipal and private voucher schools. Global enrollment rates also differ, where they are noticeably lower in schools below the cutoff. The influence of unbalanced variables does not seem to matter unless there is a jump at the threshold (which later in Table 24 I will show there is not). Nevertheless, unbalanced covariates will be included as covariates in the estimations to control their influence. The inclusion of covariates should not affect the estimates, but only increase precision. The lower section of Table 16 presents the classification rules, i.e., average scores or average proportion of students scoring above a certain score. As expected, schools

above the cutoff score higher in all the classification criteria. Differences between SIMCE scores above and below the threshold vary about a third of a standard deviation.

For the analysis of the impact of "in recovery" classification I use two rounds of schools' classification pooled together because the number of schools classified in this category is quite small. I use the classification rounds of 2012 and 2013. My analytical sample uses a bandwidth of 0.467, which reduces the sample from 5,768 observations (2,885 schools) to 1,063 observations (684 schools). Descriptive statistics of the observations is presented in Table 17.

Outcomes

I explore how the school classification affects the following sets of school level outcomes:

High stakes outcomes. High stakes outcomes are performance measures that have consequences attached. The SEP law uses the average math, language and science SIMCE scores on 4^{th} grade students to classify schools. I assess the impact of the school classification on SIMCE average scores on math, language and science. Math and language tests are similarly based on the curriculum every year these test are passed. Science, however, alternates one year based on the social sciences curriculum and one year based on the natural sciences curriculum. In year 2012 the science SIMCE tested social sciences, whereas in 2013 it tested natural sciences.

I assess the impact on high-stakes outcomes the year the school received the classification and the year after. If there is any impact the first year, then it is important to assess whether those impacts are persistent through time. If there is no impact on the first year, then it is worth assessing whether there is impact the following years, as it is possible schools need time to adjust their practices to improve their results.

Average of several outcomes above and below the ICE index threshold for the "autonomous" classification are presented in Table 18. As expected, high-stakes outcomes on 2012 and 2013

SIMCE scores in math, language and science are higher if the school is above the threshold. The gap in SIMCE scores ranges between 10 and 14 points (between a fifth and a fourth of a standard deviation).

Average high-stakes outcomes above and below the ICE cutoff for the classification of "in recovery" classification are on panel A of Table 19. Both 2012 and 2013 rounds are pooled together for the outcomes on year *t*. Outcomes of year *t+1* are only reported for the 2012 round. The gap in SIMCE scores is much smaller than for the threshold of "autonomous" schools, ranging between 4 and 7 points.

Low stakes outcomes. Following Figlio and Rouse (2006), I consider low-stakes outcomes those outcomes achieved by students on a grade that does not affect the classification status. These are SIMCE scores for grades other than 4th grade. The available data contains test scores for math, language and science for 8th graders. There is no data for year 2012 because 8th grade students were only tested every other year before 2013. After 2013, 8th grade students are tested every year.

Average of low-stakes outcomes above and below the ICE index threshold for the "autonomous" classification are presented in panel B of Table 18. Academic outcomes of a low-stakes grade are also higher above the cutoff, with SIMCE scores gaps of 11 and 12 points (a fifth of a standard deviation).

The averages of low-stakes outcomes for the "in recovery" cutoff are presented in panel B of Table 19. Because of the peculiarities of the evaluation calendar for 8th grade, there is only data on 8th grade test scores on 2013. Thus, when the average outcome is reported for a subject on

year *t*, it only considers the round of 2013; and when the reported average above and below the cutoff is for a subject on year *t+1*, then it only considers round 2012.

The third set of outcomes considers several potential mechanisms through which the incentive impacts the high-stakes outcomes:

School behavioral response. I assess the impact of accountability pressure on the number of students taking the tests, which is measured by the number of tests in the three tests they take in 4th grade. I also assess the impact on the number of students identified as having special needs. For this I created a composite measure of students with special needs which is the sum of the students with special needs enrolled in a differential group in the whole school and the number of students with special needs integrated into regular programs in the whole school.

I also assess the impact on the number of students retained in 3th grade (the grade prior to the grade when the high-stakes test is passed), the number of students that switch schools in 3th grade, and the number of students that approve 3th grade.

I consider as potential mechanisms all the criteria used in the formulas to classify schools the years following the classification. Thus I assess the impact on the proportion of students scoring above 250 and 300 SIMCE scores, school improvements assessed with the ICE index, and all its subscales: approval rates, retention rates[54], teacher evaluations, SNED improvement,

[54] School approval rates and school retention rates are for the whole school and not for 3th grade only as I assessed earlier as a potential mechanism to improve the classification the following years. The models to test approval rates and retention rates for the whole school do not include as a covariate the of 4th grade enrollment because it is also part of the denominator of the outcome which is total enrollment of the school.

SNED integration and SNED initiative (see Table 7 for descriptions). The last three indices are on a scale from 0 to 100.

The teacher evaluation index is only available for public schools, as teachers from private voucher schools are not subject to the national evaluation system of teachers. The construction of the index of teacher evaluation is constructed by the Ministry (details in Table 7). The index goes from -1 to 1, where -1 would be the score that a school gets if all the assessed teachers are classified as incompetent, 1 would be the score a school gets if all its teachers are competent or excellent, and zero represents a balance between competent or excellent teachers and incompetent teachers.

Panel A of Table 20 present the average of all of these outcomes above and below the "autonomous" cutoff. The impact of all the outcomes mentioned above is assessed for the year following the classification and the year after.

In Panel A of Table 21 I present the average of the outcomes for the "in recovery" cutoff. In the evaluation of this cutoff I assess the impact of the "in recovery" classification on school behavior, student and teacher body composition only on outcomes on year *t*. I do not assess the impact on outcomes on *t+1* because those are available only for the 2012 round of classifications and the number of observations is very small.

Student body composition. I assess the impact of school classification in the enrollment of students across schools, tuition fees and student composition. To evaluate the impact on enrollment I use three measures. One considers the overall measure of enrollment in the school. The other measures are the specific enrollment rates in grades 1^{st} and 7^{th}. I focus on these grades

because they are the grades in which we see great movement between elementary schools in Chile.

I assess the impact of school classification on students' enrollment- and monthly-fees for private voucher schools. Private voucher schools are allowed to charge fees on top of the voucher for students since 1993. However, there is a restriction for priority students. If the school adopted the SEP law, then schools cannot charge tuition to these students. The enrollment- and monthly-fees are categorical variables, where 0 means the school is free, 1 means the school charges between 0 and 15 USD, 2 means the school charges between 15 and 37 USD, 3 means the school charges between 37 and 75 USD, 4 means the school charges between 75 and 150 USD, 5 means schools charges above 150 USD (although none of the schools of the sample charged this much). I treat this outcome as continuous. This data is reported by the school. Data is not available for 2012, but it is for the year after.

I also assess whether the classification of the school may have affected the socioeconomic composition of the school the following years, specifically, students' socioeconomic level in 4th grade. The SES level of the school is a government constructed measure that classifies schools on four categories (low SES, mid low SES, mid SES, mid high and high SES). These categories are constructed from four variables: educational level of the mother, educational level of the father, monthly household income and schools' vulnerability index. The three first measures are obtained from a parental questionnaire that is passed along with the SIMCE test. The fourth measure is a government vulnerability index[55]. Contrary to previous measures of SES levels of the schools in Chile, this measure is calculated every year with annually collected data. I

[55] This index is the IVE-SINAE, calculated by JUNAEB (*Junta Nacional de Auxilio Escolar y Becas*).

transformed the categorical measure into a measure of change of SES category so that it represents changes in the SES group between the SES level used to classify the school and the SES level the years after. The variable could range from +3 if the school was a Low SES school on year *t-1* and Mid- or High-SES level on year *t*, to -3 if it happened the other way around. A zero value would mean the school did not change its SES reference group.

I also assess the impact on the number of beneficiary students. The number of beneficiary students in the school is a measure of socioeconomic composition of the school because only students from among the 40% poorest students are considered as such (a detailed description of who is considered a priority student can be found on footnote 17).

Panel B of Table 20 present the average of all of these outcomes above and below the cutoff of "autonomous" category, and on Panel B of Table 21 for the "in recovery" cutoff.

Teacher body composition. To explore the impact of school classification on the situation of teacher I use two measures. The first measure is the number of teachers teaching in the school. The second measure is the number of hours teachers are hired to teach in the school. The means of these outcomes above and below the ICE threshold for "autonomous" classification are presented on Panel C of Table 20, and on Panel C of Table 21 for the "in recovery" cutoff.

Data limitations

One of the main concerns when using school level data is the confounding influence of mean reversion (or regression to the mean). Mean reversion means that if a variable was measured substantially above or below the mean on a first measurement, it is likely that in the second measurement it will score closer to the mean. Particularly sensitive to mean reversion are small schools. All else equal, smaller schools have mean scores with higher sampling variation and

thus are more likely to have a lucky year or a very poor year (Kane & Staiger, 2002; Chay, McEwan & Urquiola, 2005). In order to address mean reversion, I take two precautions. First, I control for enrollment rates in all the models. Second, the models include the ICE index which in part is built from historical school performance (with scores up to 4 years prior to the classification the school receives on 2012).

Another limitation is that to explore potential mechanism of the law, I use administrative data. The advantage of these data is that it is readily available for almost every school. However, it is for the most part a crude measure of some indicators, and there are several interesting indicators missing. For example, I do not have access to specific data on the number of students with special needs in 4th grade. I have the number of students with special needs in all the school, but not specific to 4th grade students. So the potential mechanism of the school of identifying students as having special needs (and therefore leaving their results out of the school average) to improve the school average will be tested only with an aggregate measure for the whole school. Another example of how the administrative data can be limited is that data available on the percentage of achievement of school improvement plans (that contain the percentage of completion of actions on curriculum management, school leadership, school climate and educational resource management) is not clear on what it reports and how it was collected[56]. This

[56] For the first four years of the policy (2008-2011), the degree to which the actions of the school improvement plan were met was evaluated by a policy inspector, who then reported to the Ministry the percentage of accomplishment of the school improvement plan. The data for 2012 is not clear how it was collected. Different staff from the Ministry would provide different responses on what is assessed and how was the data collected. For 2013, the Ministry does not provide any data on the level of achievement of the actions of the PME. From 2014 onwards, the Ministry adopted a different approach in the creation, implementation and evaluation of the PME in the realm of a greater school reform in Chile.

will not allow me to assess whether there is an impact of the classification on such school processes.

There is also a specific limitation with the data available for 2012. There is no data about which schools are for-profit. Private voucher schools that are for-profit may behave in a different way than those that are not-for-profit. However, this data is available for year 2013. I perform a robustness check with data from 2013. I test this hypothesis using data for the for-profit distinction for the following year on the premise that there is not likely to be instability from year –to-year.

In the next Chapter I present the results for the estimations of receiving the classification of "autonomous" and "in recovery" separately. For each I present the arguments and tests for a specification choice and assumption check, and then I go over each one of the research questions.

Chapter 5 – Results

5.1. Impact of receiving the classification of "autonomous"

Specification choice and assumption check

The fuzzy regression discontinuity relies on the assumptions of both the instrumental variables approach and the regression discontinuity approach. The IV design generates unbiased estimates of the effect of school classification if the instrument is said to be "relevant", i.e., there is a non-zero association between instrument and treatment variable. Table 22 shows how predictive is the ICE index on whether the school is classified as "autonomous" on 2012. The table shows the first stage estimates from equation (5). Column (1) presents the estimates without covariates. The estimates in this column shows there is a significant and positive difference in the receipt of a classification of "autonomous" above and below the threshold. All the regressors are jointly statistically significant (F=2885.91, p=0.000). Column (2) adds school characteristics as covariates (i.e., type of school, SES of the school, rural/urban and enrollment rates) [57]. The addition of covariates does not affect much the estimated coefficients. This is not surprising considering the controls do not significantly change at the cutoff (explanation follows). Again, all the regressors are jointly statistically significant as well (F=2528.86; p=0.000). The estimates of columns (1) and (2) show there is a jump and a kink in the relationship above and below the threshold. The jump indicates that receiving an ICE index equal or above zero –the centered threshold score in schools' respective SES group- increases the probability of being classified as "autonomous" by 29 percent. The kink is represented by the coefficient of the interaction term that is positive and significant[58]. The jump and kink indicate that as schools have

[57] This model is similar to the one in Table 14, column (10), but it has the addition of the interaction term.

[58] The coefficient of the kink is not quite interpretable in this linear probability model as the coefficient exceeds the zero-one range of the outcome.

higher ICE indices, the probability of the school receiving a classification of "autonomous" increases. Similar estimates appear when I analyze the 2013 round of school classification and for rounds 2012 and 2013 pooled together (see Table C.2.).

Figure 3 shows the proportion of schools classified as "autonomous" on 2012 by the ICE index centered at the median of each SES group. The proportion of schools classified as "autonomous" is continuous near the cutoff on both sides, but discontinuous at the threshold. Note that on the left side of the graph the classification of schools as "autonomous" below the cutoff is zero; whereas above the cutoff the proportion of schools classified as "autonomous" increases as the ICE index of the school gets further away from the median of the school's SES reference group. This may be a result of the classification formula as schools getting high ICE indices are also more likely to comply with the other classification rules (as they refer to SIMCE scores which is the main component of the ICE index).

Another assumption of an IV is that potential outcomes are independent from the instrument, i.e., the instrument is exogenous. Whenever there is only one IV, this assumption is not testable. However, the models have more than one instrument (consider the interaction term between the running variable and the IV as second IV), thus, I test for overidentifying restrictions using the Sargan-Hansen test[59]. This tests the null hypothesis that all the instruments are valid. The statistic (x^2 (1) =0.004, p=0.9514) indicates we do not reject the null hypothesis and therefore the overidentifying restriction is valid.

[59] The Sargan-Hansen tests, tests the join null hypothesis that the instruments are uncorrelated with the error term. This involves estimating the 2SLS with one instrument A, computing the residuals of the second stage, and then assessing the correlation between the residuals and the other instrument B. If correlated, then B is not a valid instrument. This proceeding is repeated with the other instrument.

One of the assumptions of the RD is that the cutoff score(s) that determine treatment is exogenous[60]. Although the school classification may incentivize schools to raise their classification scores or decrease the SES of the school prior to the classification of the schools, it is unlikely there was manipulation of the scores around the eligibility thresholds for three reasons. First, at the time of the administration of the tests, schools had no knowledge of the median test scores of their SES group (reference threshold). Second, tests are centrally scored. Third, as mentioned earlier, classification formulas are not easily interpretable.

There is no direct way of testing whether schools can manipulate the running variable. However, McCrary (2008) proposes to plot the number of observations in bins and assess whether there is a discontinuity in the distribution of observations of the running variable at the threshold. A discontinuity would suggest there is some manipulation. If the incentive of the "autonomous" classification is really desirable, then perhaps we would find a big density of schools right above the threshold of the ICE index. Figure 4 presents the histogram. There appears to be a slightly lower density of schools with ICE index above the threshold of their SES group. McCrary (2008) density test[61] suggests there might be weak evidence of a change in the density of the running variable on either side of the discontinuity (log difference in height: -0.14; p=.08). Because the running variable in this test is stacking all the schools' ICE scores by SES centered at their specific thresholds, it could be hiding different density shapes in different SES groups. To be sure there is no manipulation I analyze the density of the running variable for the

[60] Any evidence of manipulation of the running scores near the cutoff would question the RD design (Urquiola & Verhoogen, 2009).

[61] McCrary's (2008) density test, tests the null hypothesis of continuity of the density of the running variable at the threshold. It entails two steps. In the first step it partitions the running variable into bins (where no bin includes points on both sides of the threshold) and calculates the number of observations per bin. The second step consists of a weighted local linear regression at each side of the threshold where the height of the bins is regressed on the midpoints of the bins. Then the parameter of interest is the log difference of the coefficients on the intercepts.

schools in different SES groups separately. In Figure C.1. I show the estimated densities by SES group separately. Although the figures for schools of mid-low SES and mid-SES show some discontinuities at the threshold favoring the lower SES, they are not statistically significant. In principle then, the regression discontinuity analysis conditional on school's SES should reduce any effect of changes in SES levels.

Another assumption of the RD is that the relationship between the outcome and running variable is modeled correctly. Two concerns arise on this matter. One is that the choice of bandwidth influences the results[62]. I use a bandwidth of 0.706 points on the ICE index from each side of the median of each SES group[63] [64]. As a specification check, in the different columns of Table 23, I present the estimates run on samples with different bandwidths, i.e., including schools as far from the threshold as 1.2 points above the median of the SES group of reference and as close to the eligibility threshold as 0.5 points. When we compare the models horizontally in this table, the estimates illustrate the trade-off between precision and comparability. As the bandwidth of the sample narrows, the standard errors tend to grow. As the sample expands further away from the cutoff, the less comparable are the groups above and below the cutoff. However, the estimated effects do not differ much across the columns. A second concern is the polynomial choice. I use two strategies to select the polynomial order. First, as suggested by Lee

[62] The standard approach is to choose a bandwidth that minimizes the mean squared error of the RD point estimator (Skovron & Titiunik, 2015). This depends on the density, variance and curvature of the data near the cutoff (Scott-Clayton, 2008).

[63] Following Imbens and Kalyanaraman (2012) optimal bandwidth calculations of the first stage leads to bandwidths close to 0.3 for several outcomes, bandwidth that is too small to generate estimates. Following Calonico, Cattaneo and Titiunik (2014), the bandwidth for several high-stakes outcomes are close to the bandwidth estimated for math test scores on year *t* of 0.706 (plus/minus 0.1). For simplicity, I use this bandwidth for all outcomes of the cohort of 2012. I test whether the results are sensitive to the size of the bandwidth. Results do not vary much.

[64] I use triangular weights to weight more heavily those schools that are closer to the threshold than those that are further away.

and Lemieux (2009), I use the Akaike information criterion (AIC) of model selection[65]. Second, I assess the significance of the F-statistic as I add higher order polynomials into the models. I use a linear model[66] as it presents the lowest AIC statistic, and adding higher order polynomials present F-statistics that are not significant. As a specification check in the different rows of Table 23, I report a number of specifications to illustrate the robustness of the results.

A third assumption is that there are no other factors confounded with the forcing variable (Schochet et al., 2010). This implies no other covariates should present a discontinuity at the threshold. This assumption is checked by examining the continuity of observable pre-treatment variables that could be related to potential outcomes. Table 24 presents the estimates of β_1 from Equation 6 for several outcomes which are pre-treatment variables. None of them present a jump in the threshold.

What is the impact of accountability pressure on high-stakes outcomes?

I begin with the analysis of the impact of being classified as "autonomous" on high-stakes outcomes. Before proceeding to the regression results, it is useful to examine graphical representations of the effects of school classification. Figure 5 plots high-stakes outcomes, specifically the average math (panels A and D), language (panels B and E), and science (C and F) SIMCE scores on year *t* and *t+1* by the ICE scores in reference to the median of the SES group of the school (represented by the vertical line). On either side of the threshold I added a line representing a local linear regression but without any covariates. Overall, the six panels show a positive association between test scores and the ICE score. However, none of the panels

[65] The Akaike information criterion measures the relative goodness of fit of a model. Is the estimated residual variance for the model (Jacob, Zhu, Somers & Bloom, 2012).

[66] According to Angrist and Pischke (2009), as the sample gets smaller by trimming it closer to the threshold, the number of polynomials needed for the model should go down. This same suggestion is provided by Gelman and Imbens (2014).

show discontinuities at the threshold and no significant changes in the slopes of the linear predictions on the two sides of the threshold.

Before proceeding to the regression results I will explain the main format of the following tables. Tables 25 through 29 follow the same format. Each row focuses on a different outcome, with each cell containing the estimated parameter β_1 of Equation 6, the robust standard error on parenthesis and the sample size. The first column contains the linear model presented in Equation 6 without covariates. Covariates are included in the second column. The covariates included are the type of school, whether the school is urban or not, enrollment rates in 4th grade and the SES level of the school. The first two columns consider the full sample, whereas any other added column to the right considers sub-samples detailed by the column heading. All the models in the columns that contain subsample analysis control for covariates. The models for all outcomes were calculated using OLS.

Table 25 lists a set of high-stakes outcomes for the year the school received the classification of "autonomous" (*t*) and a year after (*t+1*)[67]. The models in columns (1) show positive effects of receiving a classification of "autonomous" on math, language and science test scores, both the year of the classification and one year later. In column (2) I add covariates. The inclusion of covariates does not change the point estimates significantly, but decrease the standard errors. Point estimates for schools on the margin of the threshold range between 0.25 to 2.72, which is an effect size between 0.005 and 0.05. However, the effect is never statistically different from zero (with or without covariates). I found that the smallest impact that would be statistically

[67] Schools receive the 2012 classifications on November 2011. The academic year for 2012 goes from March to December. On October 2012 SIMCE tests are passed to 4th grade students.

significant in this regression discontinuity analysis ranged from 0.12 to 0.16 (see Panel A in Table C.6.).

I performed robustness checks with other cohorts of data. In Table C.3., columns (1) and (2) in the upper panel, we find similar results for the cohort of 2013. Although the estimates are sometimes positive and other negative for high-stakes outcomes the year the school received the classification, the effect is not statistically different from zero. To further increase power I pooled together the cohort of 2012 and 2013 and added fixed effects per year to account per any cohort trend on SIMCE scores. The resulting estimates are presented in Table C.4. The effects for all SIMCE scores the same year the school received the classification are positive, with point estimates ranging from 0.18 to 1.69. These estimates are not statistically different from zero[68].

The lack of evidence of impact on academic outcomes is consistent with studies of accountability pressure where, overall, schools receiving good accountability grades are not responsive to accountability pressure (Rockoff & Turner, 2008; Weiner, Donaldson & Dougherty, 2016). This apparent lack of response is quite different from schools receiving failing grades, where the literature shows threatened schools seem to positively respond to those grades.

[68] I also did the exercise of estimating the models using as an instrumental variable the other rating scores that determine the school classification. Of course the estimates of these models affect schools that are close to those specific frontiers, and therefore they are not directly comparable. Nevertheless, if there was a frontier that for repeated years showed an impact on outcomes, then we could think that there is a specific subsample of schools (defined by those induced by the compliance of that specific rating score to be classified as "autonomous") for which accountability pressure has an effect. The results of this exercise are presented in Table C.5. On the top of each column the rating score used as IV is displayed. Each cell contains the coefficient of interest of the second stage equation for all the outcomes listed in each row. The bandwidth used for the estimation of each IV is listed at the bottom of the column. The estimates suggest none of the types of rating scores consistently generates an impact on the outcomes. Only the models that use the IV of column (6), which is whether the proportion of students scoring above 250 points is above the median of the schools' SES group on year t-4 (which corresponds to year 2008), shows impacts on the outcomes. The other IV's that also use to the proportion of students scoring above 250, but for other years, shows no impact on high-stakes outcomes.

Different responses may be driven by the stakes associated with the grades, i.e., by how much schools value the associated rewards or how much they fear the associated sanctions.

The lack of evidence of impact of receiving a good grade on academic outcomes is similar to the evidence found for another incentive program set in place in Chile. The SNED program is a collective incentive for teachers if the school where they work achieves good results in comparison to schools with similar SES -similar to how schools are classified in the SEP accountability scheme-[69]. An evaluation of this program by Rau and Contreras (2009) shows schools winning the SNED recognition does not have a consistent impact on schools' academic outcomes; however there is some impact on one of the tested cohorts[70].

Heterogeneous effects of receiving the classification of "autonomous"

I examine not only the local average effect of accountability pressure, but also explore potential heterogeneous effects for groups where there is some suspicion they may react differently to the pressure. First, I assess whether the effect of school classification differs depending on the classification the school received the year prior. Schools that are just below the classification threshold the year before may have more incentives to perform better than the schools that are just above the threshold. Second, I divide schools according to the average socioeconomic levels of the families. There is evidence that the introduction of the SEP law had a greater impact on

[69] The SNED and the SEP incentive schemes for high achieving schools differ in that (i) the beneficiary of the reward in the case of the SNED are individual teachers whereas in the SEP, the beneficiary is the school as an organization, (ii) the SNED's incentive is offering an annual bonus to teachers of almost 70 percent of a monthly salary, and the SEP does not offer more resources to the school, but more autonomy to allocate those resources, (iii) schools are also classified as 'winning' the SNED or receiving the classification of "autonomous" differently.

[70] Rau and Contreras (2009) use a regression discontinuity approach to assess the impact of winning a SNED award. They found a positive and significant impact only in one of the five cohorts tested (2005/2006).

schools that enroll students from lower SES; for schools that enroll students from higher SES the impact was negligible (Mizala & Torche, 2013).

Third, I assess whether private voucher schools that are for-profit behave differently than private voucher schools that are not-for-profit when they face accountability pressure. Schools that are for profit may face different incentives than non-for-profit schools (Peterson, 1974; Fernandez, 2009); and there is evidence in Chile that for-profit and non-for-profit schools impact differently the outcomes of the students study (Zubizarreta, Paredes & Rosenbaum, 2014)[71]. For example, for-profit schools may have stronger incentives to improve their test scores, as being classified as "autonomous" gives schools more discretion to allocate resources. Although all extra resources have to be reported as being allocated in activities associated with curriculum management, leadership, school climate and resource management, the law does not forbids schools to substitute other funds allocated into those areas. This means schools could substitute the funding source for those areas. For-profit schools could then be able to make more profits if they receive the classification of "autonomous" rather than "emergent" or "in recovery".

Finally, I assess whether the effect of school classification differs depending on the classification other schools receive within the same municipality. Some suggest that accountability systems may be more effective in education markets that are least competitive (Deming & Figlio, 2016). However, this may not be the case in a context where there is school choice. There is evidence that in such contexts the impact of school classification may be

[71] The authors compare the academic outcomes of students that switch from a public school to a private voucher school for-profit to those students that switched to a private voucher school not-for-profit. They use SIMCE data from Santiago of students who attended 8th grade on the public school and their SIMCE scores on 10th grade in the new private voucher school. They found students that switched to a private voucher school that is not-for-profit performed 0.1 and 0.2 standard deviations (on language and math tests respectively) above the group of students that attended a for-profit school.

moderated by the level of competence among the schools, with higher impacts on more competitive markets (Weiner, Donaldson & Dougherty, 2016).

To examine heterogeneity in the effect of receiving a classification of "autonomous" I run separate regressions for different subsamples. Table 25 in columns (3) and (4) show the estimates of two subsamples of schools, those that received an "autonomous" classification the year prior, and those that did not receive a classification of "autonomous" the prior year, as these two sets of schools may face different incentives to improve. The estimated coefficients on both these samples are again positive, but not statistically significant.

In Table 25, columns (5) to (8) restrict the sample to homogeneous groups by SES. Overall none of the estimates is statistically different from zero. However, it is worth looking at the signs of the effects. High SES schools show consistently negative coefficients. When looking at the impact on math test scores the year of the classification, the point estimate suggests that on average, schools that received a classification of "autonomous" achieved five points less than those schools that just missed the classification of "autonomous". Similar results suggest the estimates of language and science test scores. The magnitude of the effect seems to slightly decrease for the following year in all the subject areas. These results could indicate high SES schools take a rest once they are classified as "autonomous" or perhaps that there is some regression to the mean. The impact on mid-low SES and mid-SES subsamples seems to be consistently positive, although quite small, and again nothing statistically different from zero. The impact on low SES schools is less clear, as estimates are in some tests positive and others negative.

Finally, the columns (9) and (10) of Table 25 show the estimates for schools in highly competitive municipalities and in less competitive municipalities. As Weiner, Donaldson and Dougherty (2016) suggest, schools in more competitive local contexts may put more effort in earning a distinction. These distinctions may help schools attract more students to enroll. Following Weiner and colleagues (2016), I consider a municipality to be highly competitive if the share of schools classified as "autonomous" in the municipality is equal or greater than 0.2[72]. The estimates suggest the impact of receiving the classification of "autonomous" in a highly competitive context is small and positive, but not statistically different from zero. The impact of an "autonomous" classification on a less competitive context is close to zero, but again not statistically significant. These findings are contrary to what Weiner and colleagues (2016) found for schools on highly competitive contexts in Rhode Island, where schools that just missed the good qualification increased their test scores.

Using the 2013 cohort, I also estimate the impact of school classification on high-stakes outcomes on private voucher schools that are for-profit and those that are not-for-profit. The evidence is presented on Table C.3. in columns (3) and (4). The estimates indicate the response of both types of schools to accountability pressure is not statistically different from zero.

What is the impact of accountability pressure on low-stakes outcomes?

I next assess whether there is any impact on low-stakes grades. I consider 8th grade as a low-stakes grade because there are no consequences for the school enacted by the government. This does not mean that the achievement of the school in 8th grade may not be considered as high-

[72] Figure C.2. presents a scatterplot of the share of schools in a municipality that receive the classification of "autonomous". The horizontal line divides the plot between those municipalities that have a share of 20% or more of schools classified as "autonomous". The plot shows there is enough variation of the share of schools classified as "autonomous" within the municipalities.

stakes for other purposes. In fact, 8th grade achievement may be high-stakes for parental decision to choose schools. Furthermore, it seems likely that any impact on 4th grade outcomes may, in the long term, be reflected in 8th grade test scores, as student learning is more or less cumulative (Gilraine, 2016).

Table 26 presents the estimates of the impact of accountability pressure on math, language and science test scores on 8th grade, in all the schools within the 0.706 bandwidth. The impact of being classified as "autonomous" on math, language and science tests scores the year after the classification[73] is no different from zero. The minimum detectable effect size ranged between 0.14 and 0.15 (see Panel B, Table C.6.). As robustness check the impact of accountability pressure on low-stakes outcomes using data from the 2013 cohort. The estimates are presented in the lower panel of Table C.3. The estimates in column (1) show there are some positive and significant effects favoring schools classified as "autonomous", those differences disappear when I control for covariates in column (2).

These results are consistent with the theory of incentives, as it predicts that schools may put effort on rewarded outcomes and not in other outcomes unless they are complementary to the rewarded outcomes (Holmstrom & Milgrom, 1991). In comparison with the accumulated empirical evidence where mixed results are found, the results found in this study are not surprising.

If school accountability increases incentivized or non-incentivized outcomes, what mechanisms drive those improvements?

[73] 8th grade was not tested on 2012, and therefore there are no low-stakes test scores for year t.

Making a mechanism evaluation could seem irrelevant if the previous results have shown the policy had no impact on the expected outcomes. However, understanding the policy's mechanisms can help better understand the impacts of its design or the moderators of its impact (Ludwig, Kling & Mullainathan, 2011).

The accountability policy in Chile had clear mechanisms through which it wanted to affect the outcomes. The spirit of the policy was that schools would improve their results by defining and implementing clear guidelines and actions to improve curriculum management, school leadership, school climate and educational resource management. Data on these areas is not available. However, as discussed earlier in Chapter 2, the empirical evidence on other countries shows there are several other potential mechanisms through which schools can improve their outcomes. In this section I explore several of those mechanisms. Because I have already shown there is no impact on high- or low-stakes test scores, the purpose of this section is not to explain why those test scores could have significantly improved. But the purpose of this section is to assess whether schools facing accountability pressure may have activated some mechanisms to improve the high-stakes test scores (even if that did not have an impact on test scores).

There are several changes schools can make to deal with the accountability pressure. They can improve school quality as they improve the performance measure, or they can simply affect the performance measure without really impacting learning. Some of the strategies identified in the literature in Chapter 2 include removing low-achieving students from the pool of test takers, narrowing of the curriculum, focusing on marginal students, and increasing instructional expenditures. There is also some evidence of teacher and student mobility across schools. In this section I perform an exploratory analysis of some of these mechanisms.

School behavioral response. Under accountability pressure schools have an incentive to prevent low-achieving students to take the tests or to classify more students as having special needs so that their tests scores do not count against the school. I assess whether the classification of the school affects the number of students that take the SIMCE test in 4th grade. If schools that just missed the cutoff were trying to leave low performing students without tests, then we would find a positive coefficient favoring the schools classified as "autonomous". The first two rows of Table 27 show the actual estimates of the impact of receiving a classification of "autonomous" the year of the classification and the year after on the number of students taking the tests in 4th grade. Column (1) shows the estimated effects without controlling for covariates, column (2) shows the effects controlling for covariates. The difference in outcomes between schools receiving the "autonomous" classification and the other classifications is positive as expected, but not statistically significant. When assessing whether there is any differential impact due to the different levels of pressure the school may face by the classification the school got the previous year, shown in columns (3) and (4), there also does not seem to be a significant difference between schools receiving the classification of "autonomous", and the other schools.

I also assess whether there is an impact on the number of students considered to have special needs. If schools right below the cutoff wanted to remove from the pool of tested students those with special needs, then we would expect to find a negative coefficient for the "autonomous" schools. Rows 3 and 4 on Table 27 show the estimated effects of this measure. The effect displayed in columns (1) and (2) shows the number of students classified as having special needs varies from differences that are one year negative and one year positive and nothing consistently and statistically different from zero. However, these findings should be considered carefully, as the number of students with special needs is measured for the whole school and not 4th grade in

particular. Therefore, it is possible the overall measure hides imbalances in different grades. The estimates in columns (3) and (4) also show there is no impact on the number of students classified as having special needs between schools above and below the threshold facing different pressure.

Just as schools may have an incentive to prevent low-achieving students to take the test or to classify more students as having special needs, schools may also have an incentive to retain low-achieving students the year prior to the grade the high-stakes test is applied. Therefore I assess the impact of school classification on the number of students retained in 3th grade. If schools that just missed the "autonomous" classification were retaining more students in third grade, then we would expect to find a negative coefficient for the "autonomous" schools. Rows 5 and 6 in Table 27 show the estimated effects. The estimated coefficients are negative but not statistically significant. Because some low-achieving students may not be retained in 3th grade, but may drop-out instead, I also assess the impact of the classification on the number of students approved in 3th grade. This measure represents the total enrollment in 3th grade minus the retained students and those that dropped out. If schools that just missed the classification threshold were trying to achieve their test scores by pushing away low-achieving students or retaining them in 3th grade, then we may see that they decreased the number of approved students. Thus, we would expect a positive and significant effect for the "autonomous" schools. However, the estimated effects shown in rows 7 and 8 in Table 27 show a negative impact, but not statistically different from zero.

Following the same rationale as explained earlier, schools may face the pressure of pushing out of high-stakes grades low-achieving students. This may be reflected in a greater number of students switching schools the year prior to 4^{th} grade, altering the distribution of students across

schools. I assess whether there are differences between schools classified as “autonomous” and the other schools regarding the number of students that switch schools in 3th grade the year the school received the classification and the year after. Rows 9 and 10 of Table 27 show the estimated impacts. None of the differences are significantly different from zero.

Under accountability pressure schools also have a strong incentive of allocating their efforts on students who are near proficiency thresholds. I assess the impact of school classification on the percentage of students scoring above 250 and 300 test scores relative to the median of the school’s SES group. This outcome may also be considered as high-stakes considering it is also a determinant of the classification the schools get. A school that just missed the classification of “autonomous” may try to get such classification the following year by increasing the percentage of students scoring above 250 or 300 SIMCE points. Rows 11 to 14 of Table 27 show the impact of accountability pressure on the percentage of students scoring above 250 and 300 test scores relative to the median of the school’s SES group. The estimates show the difference between school above and below the threshold are not statistically different from zero, for all schools and differentially for schools that may face more or less pressure given their prior year classification.

Under pressure, schools may also work in improving school practices. I assess the impact of school classification on several indicators of school improvement. One measure of school improvement is the ICE index. This measure should be considered carefully, because it is also a high-stakes measure, as it determines the classification the school gets the following year. Schools may have the incentive to improve the ICE index without actually improving the school practices. This could be particularly the case for one sub-scale of the index that is not a measure of the outcomes of the school (such as SIMCE scores, approval rates, retention rates), but a measure of the perception of school processes (such as the SNED initiative subscale), which

come from the responses of the 'school principal' to a survey. Thus, I assess the impact of the classifications of the school in all the sub-scales of the ICE index separately. Results are in Table 27, rows 15 to 26. Receiving a classification of "autonomous" does not affect the ICE scores significantly, and none of the sub-scores of the ICE index are significantly different from zero.

Lastly, I assess the subscale of teacher evaluations which is an index of the balance between teachers that where considered excellent or competent and those that are considered incompetent by the teacher evaluation system. Note that the sample size of the models assessing these outcomes decreases around a 40%. This happens because only public schools have teacher evaluations. The estimated effect of receiving the classification of "autonomous" on teacher evaluations is in the last two rows of Table 27. The effect is not different from zero.

Student body composition. In an accountability system the classifications or grades schools get can affect the way parents choose schools for their children and it also may affect the way schools choose their students. There are three ways in which parents can respond (Hart & Figlio, 2015). First, they could not respond to new information about the quality of the school. Second, new information about the quality of the schools may affect the choice of parents from low-SES backgrounds who did not have much information prior to the provision of this new information. Third, parents from high-SES backgrounds could have more capacity to respond to the provision of new information.

There are also different ways in which a school can react to accountability pressure. It is possible schools receiving better qualifications are in higher demand. This could mean the school enrolls more students, or gets more selective in terms of the academic background of incoming students, or starts charging more fees, or a combination of all these. In the particular case of the Chilean context, it is possible that schools who are receiving lower classifications enroll students

with lower SES background, as this will potentially increase their classification for the following year (because the classification formulas are relative to the median of schools with similar SES).

Table 28 shows the estimates of the impact of receiving a classification of "autonomous" on a series of outcomes that reflect student mobility between schools. The first two rows of the table shows the impact on the total enrollment of the school the year the school was classified, and the year after. Changes in enrollment levels may mirror both changes in parental decisions to choose schools as well as differences in schools' capacity to attract students. Column (2) shows there is a minor positive difference favoring "autonomous" schools; however this difference is not statistically significant. Now if we look at columns (3) and (4) of Table 28, I estimated the impact separately for schools were "autonomous" the previous year and those that were not. The estimates indicate that schools that face more pressure because they were not "autonomous" the previous year have a positive impact on the number of students enrolled, whereas schools that face less pressure have a negative impact. Nonetheless, none of these estimates is statistically significant. This pattern is also present in enrollment in 1st and 7th grade, which are the grades in primary education where we see greater movement of students between schools.

Schools receiving good classifications may become more selective by increasing their price[74] or by becoming more selective in terms of the socioeconomic level of their students. I assess the impact of receiving a classification of "autonomous" on enrollment- and monthly-fees for the year after the school received the classification (there is no data available for the fees charged on

[74] Schools in the SEP law are not allowed to charge fees to priority students. Therefore, it is possible schools increase the enrollment- and monthly-fees to all the other students not only to be more selective, but to compensate for such restriction.

2012). Only private voucher schools are allowed to ask parents for additional fees since 1993. For this reason I restrict the sample to only private voucher schools. The impact of school classification on these outcomes is estimated using OLS, and the estimates are presented in rows 9 and 10 of Table 28. I find no evidence of differences in the amount of enrollment- or monthly-fees charged to parents a year after the school received the classification of "autonomous".

The four bottom lines of Table 28 show the impact of receiving a classification of "autonomous" on the socioeconomic composition of the schools, as measured by changes on students' SES at 4^{th} grade and the number of beneficiary students on the whole school. Row 11 and 12 from Table 28 shows the impact of receiving a classification of "autonomous" on changes in the SES level of the school the year the school is classified as "autonomous" and changes towards the year after. These models do not control for the baseline level of SES of the school to avoid a spurious correlation. The point estimates are negative the year of the classification and positive the year right after, however they are not statistically different from zero.

The last two rows of Table 28 show the effect of receiving the classification of "autonomous" on the number of beneficiary students. There is no evidence of changes in the number of beneficiary students the year the school gets the classification of "autonomous" or the year after.

Overall, the evidence shows there is no evidence of impact on student mobility between schools and composition of the schools. These results are consistent with the recent work of Mizala and Urquiola (2013), who show that rewarding schools for their good performance with the SNED program has no impact on subsequent schools' market outcomes such as enrollment rates and tuition.

Teacher body composition. The way teachers move between schools can be affected by the classification schools get. These movements can be driven by changes in the demand for teachers and/or changes in the supply of teachers. Low performing schools may want to hire more teachers or the same teachers for more hours. Teachers may prefer to migrate to high performing schools as there they may be required to put less effort into teaching.

Table 29 presents estimates of the impact of receiving a classification of "autonomous" on the number of teachers who are teaching in the school the year the school receives the classification and the year after, and also the number of hours those teachers are teaching. I find no evidence that receiving a classification of 'autonomous" is associated with differences in the number of teachers teaching in a school, nor in the number of hours taught. Nevertheless, this exploratory analysis should be considered carefully. These only assess the quantity of teachers, but not their quality. There are several other ways in which schools can respond to school classification regarding their teachers. For example, low performing schools could have decided to offer more teacher training without altering the number of hours teachers are hired.

While the literature mentioned above indicates that schools and families may respond in several different ways to incentives, I find no evidence of any type of reaction to schools receiving a good classification. However, this exploratory analysis of potential mechanisms should be considered carefully. The administrative data used is very gross. Educational improvement may require more sensitive strategies not captured by these administrative data.

In the following section I present the results of schools receiving a classification of "in recovery".

5.2. Impact of receiving the classification of "in recovery"

Specification choice and assumption check

Following the same rationale as in the previous section, here I test assumptions of the instrumental variable and regression discontinuity approaches. I first start with the assumptions of the IV design. Table 30 presents evidence that the instrument has a strong association with the treatment variable. The table shows estimates for equation 7 estimated with a linear probability model. Column (1) shows the estimates without controlling for covariates. When covariates are added on column (2) the estimates do not change much. The estimates indicate that having an ICE index below zero –the centered cutoff score in schools' in 2012 and 2013- increases the probability of being classified as "in recovery" by 0.9. This is consistent with Figure 6 showing the ICE index generates almost a sharp discontinuity at the threshold. The exogeneity of the instrument is tested using the Sargan-Hansen test. The statistic (x^2 (1) =0.085, p=0.7703) indicates we do not reject the null hypothesis and therefore the overidentifying restriction is valid.

Some of the assumptions of the RD design are that the cutoff score that determines the treatment is exogenous, that there are no confounded factors with the running variable, and that the relationship between the running variable and the outcomes is modelled correctly. There are several reasons to believe there is no manipulation of the running variable, i.e., that the treatment is exogenous. First, 2012 is the first year in which schools were classified as "in recovery", so schools may have just a broad sense of the relative position they have from the cutoff, but not an exact knowledge of their position. It is also not likely that schools manipulated the running score for the 2013 classification, because the classification considers multiple scores collected before the school received the classification of 2012. Second, tests are centrally scores. Third, unless the schools were able to see the technical reports about how schools are classified (which are only

available through the Transparency Law requests), they did not know how the ICE was index used to determine the "in recovery" classification. To complement these arguments, Figure 7 shows the density of schools around the threshold of the ICE index. There is no evidence of a discontinuity that would suggest some manipulation.

To assess whether the relationship between the running variable and the outcomes is modelled correctly I follow the same procedures as in the previous section. I perform robustness checks for bandwidth selection and for polynomial choice. Table 31 present alternative specifications for the outcome of test scores at the end of the year the school received the classifications. Different columns present samples with different bandwidths, as narrow as ±0.3, and as width as ±1. Different rows show the results for models with the addition of higher order polynomials of the running variable. The cells present the estimated coefficient ϑ_1 of Equation 8 and its robust standard errors. The cells also contain the *F*-statistic (and corresponding *p*-value) for the addition of higher order polynomials and AIC to test for model selection. Overall, the estimated coefficients do not differ much horizontally or vertically, and almost none of them are statistically significant. The first cell in column (1) shows a slightly significant impact, but this seems to be drawn by the undue influence of observations far away from the cutoff (as seen in the left side of Figure 6). Regarding the order of the polynomial, I decide to use a linear model as it presents the lowest AIC statistic for all the samples with different bandwidths, and adding higher order polynomials have F-statistics that are not statistically significant for any of the samples.

I test whether there could be any other factor confounded with the forcing variable, which implies that no covariate should jump at the threshold. In Table 32 I present estimates of ϑ_1 from

Equation 8 pre-treatment variables. There is no evidence that any of the pre-treatment variables present a discontinuity at the threshold.

What is the impact of accountability pressure on high-stakes outcomes?

Figure 8 plot the raw means of schools' SIMCE scores on math, language and science at the end of the year the schools received the school classification by the ICE index, upon linear predictions (without any covariate). Overall the three panels show a positive association between test scores and the ICE index. The linear prediction of Panel A shows a small jump at the threshold; however, the dispersion of the observations is quite large. The other two panels do not show jumps at the thresholds or changes in the slopes of the linear predictions.

The regression results are presented in Table 33. This table and the following have similar formats. Each row focuses on a different outcome. Within each cell is the estimated parameter ϑ_1 from Equation 8, the robust standard error and the sample size. For outcomes in year *t*, the models include data from years 2012 and 2013 with a fixed effect per year. For outcomes in year *t+1*, the models include only data from the 2012. The models for all outcomes are calculated using OLS.

Table 33 shows the estimates of the impact of receiving the classification of "in recovery" on math, language and science test scores in 4th grade. Column (1) shows baseline estimates, whereas column (2) adds covariates. The estimated coefficients are positive for the year of the classification, ranging between 4.75 and 1.52. However, none of these estimates are significantly different from zero. The estimates for test scores the year after the school received the classification are larger; however, the standard errors are also larger. There are no effects

statistically different from zero. The minimum detectable effect size values varied between 0.13 and 0.21 (see Panel A, Table C.7.).

This evidence is not consistent with what has been found on the literature of accountability pressure, where schools classified as low achieving show a positive and significant response reflected on high-stakes outcomes (Figlio & Rouse, 2006; Rockoff & Turner, 2008; Chiang, 2009; Allen & Burgess, 2012; Deming et al., 2013; Hussain, 2015). However, it seems such evidence is a result of more radical treatments than receiving the "in recovery" classification in Chile. Here are some examples. In New York City and Texas, low performing schools face the threat of leadership change and/or possible closure of the schools (Rockoff & Turner, 2008; Deming et al., 2013). In England, low performing schools are subject to the stigma of the public judgement and more intense interventions like changes in the leadership team, school governing board, as well as increased oversight from the inspectors (Allen & Burgess, 2012; Hussain, 2015). These same threats built into the accountability policies from New York City, Texas and England were designed into the SEP law in Chile; however the actual enactment of these threats was postponed by the creation of the Quality Agency of Education until 2020.

Another example comes from Florida. Low performing schools in Florida not only face the stigma of receiving a low classification, and the fear of replacement of the 'school principal', but also faced the fear that students may leave as they were offered vouchers to attend other schools (Figlio & Rouse, 2006; Chiang, 2009). In Chile, low performing schools also face the threat of students leaving because we also have a voucher system. However, the difference between Florida and Chile is on how widely available was the information on schools' classifications to the parents. In Florida, the grading system is quite simple to understand, and the grades schools

received were widely published. In Chile the information on school classifications is not clear what it means and is not widely available to parents.

What is the impact of accountability pressure on low-stakes outcomes?

I next assess whether receiving the classification of "in recovery" affects academic performance of grades which results are not attached to consequences in this accountability policy. These are the results of 8th grade on the SIMCE test. Due to the peculiarities of the evaluation calendar for 8th grade there is no data on their test scores on 2012. For this reason I do not use the 2012 and 2013 cross-sections pooled together. I assess the impact of receiving the classification of "in recovery" on 2012 on the outcomes of 2013, and the impact of receiving the classification of "in recovery" on 2013 on the outcomes of the same year.

Table 34 presents the estimates of ϑ_1 of Equation 8. The estimates of receiving a classification of "in recovery" on 2012 indicate there is some positive and significant impact on 8th grade test scores on math the year after the classification. The point estimate for schools on the margin of the threshold is 7.04, which is equivalent to an effect size of 0.14. The impact on other subject scores is always positive, but not statistically different from zero. The minimum detectable effect size values varied between 0.12 and 0.21 (see Panel B, Table C.7.).

The estimates of the impact of receiving the "in recovery" classification in 2013 on test scores on that same year are presented on columns (3) and (4) of Table 34. All the estimated coefficients are negative; however none of them is statistically significant.

Overall, these estimates are consistent with the theory of incentives, as effort may not be put into outcomes that are not rewarded.

Regarding the small impact found on 8th grade math scores should be considered carefully. The presence of one significant coefficient among hundreds estimated does not necessarily mean there is a relationship between accountability pressure and school outcomes. The criterion of 95% confidence suggests that if the relationship between accountability pressure and school outcomes was truly unrelated, we would still expect to find a 5% of the estimated coefficients to be statistically significant due to chance alone. The coefficient found to be significant most likely is within this 5% of chance.

If school accountability increases incentivized or non-incentivized outcomes, what mechanisms drive those improvements?

Following the same rationale as in the previous section, I explore potential impact on school behavior, student and teacher mobility across schools. The same hypothesis and explanations of the outcomes apply. I only assess the impact on outcomes the year of the school classification because I have data for the outcomes of both the round of 2012 and 2013.

School behavioral response. Schools classified as "in recovery" may find more pressure to prevent low achieving students to take the 4th grade SIMCE test. If that was the case, then we would expect a positive coefficient favoring schools "in recovery" in the number of students classified as having special needs, the number of students retained in 3th grade, the number of students switching schools in 3th grade, and a negative coefficient in the number of test taken and the number of students approved in 3th grade. In Table 35 I show the estimated coefficients. For all of the outcomes the sign of the coefficients has the expected sign; however, none of them are statistically different from zero.

Under pressure, low achieving schools may face pressure to focus on the students who are near the proficiency thresholds as defined by the classification rules. In the SEP law, not only average scores matters, but also the proportion of students scoring above 250 and 300 points in 4th grade SIMCE. Rows 7 and 8 in Table 35 show the estimated impact of receiving the classification of "in recovery" on these outcomes. There is no evidence of a significant difference between schools "in recovery" and schools receiving other classifications.

Low performing schools may also consider focusing on increasing the scores in the classification rules so that would help them classify in a higher category the following years. I assess the impact on the ICE index and all it sub-scores: SNED initiative, SNED improvement, SNED integration, approval rates, retention rates and teacher evaluations. Table 35 in rows 8 to 14 show no evidence that receiving the classification of "in recovery" impacted any of the subscales the year of the classification. There is some statistically significant impact on retention rates; however, the difference is not very meaningful in practice.

Student body composition. Schools may see the mobility of the students affected by the classification they get. Table 36 shows estimates of the impact of receiving a classification of "in recovery" on total school enrollment, and 1st and 7th grade enrollment. The impact on total enrollment is positive, and negative in 1st and 7th grade. However, none of the coefficients is statistically different from zero.

I also assess whether the socioeconomic composition of the school may be altered by the classification of the school. The last two rows of Table 36 show "in recovery" schools slightly decrease the SES level of the school and slight increase the number of poor (beneficiary) students enrolled. None of the coefficients are statistically significant.

Teacher body composition. School classified "in recovery" may invest in more teachers or in hiring the same teachers for more hours to improve the results of the students. Table 37 presents estimates for the impact on those outcomes. The evidence shows there is a positive impact on the number of teachers teaching, and the number of hours they are hired to teach. However, the impacts are not statistically significant.

Overall, I find no evidence of schools modifying their behavior, composition of their students or teachers to improve their results. The same caveats expressed in the previous section apply.

In the following chapter I summarize and discuss the results.

Chapter 6 - Summary and discussion

Summary

In the last two decades, school accountability policies seem to be the leading proposed solution to improve school quality. The idea is that the presence of standards, test-based information, and performance-based consequences will generate a strong incentive for schools to improve their performance. However, the accumulated evidence of the effectiveness of school accountability policies is discouraging. The effectiveness evidence shows a positive but mild impact on students' high-stakes outcomes, with effect sizes ranging from 0.04 to 0.33. The impact on low-stakes outcomes is not conclusive, suggesting school accountability may not be impacting true learning. Studies about the mechanisms that drive the impact on high-stakes outcomes mostly focus on changes in the school behavioral response. Fewer studies have examined teacher and student mobility across schools as an effect of the school accountability programs. The evidence from other studies show the increases in high-stakes outcomes may be due to non-desired behavior of the schools such as narrowing of the curriculum, manipulation of the pool of test-takers, increased student retention, increased number of special education placements, and focus on marginal students.

A major gap in our understanding of school accountability concerns whether there is any impact on students' performance in a context where parents are free to choose schools for their offspring, and what mechanisms drive those impacts (if any). This dissertation addressed some of these concerns. I assessed the impact of accountability pressure in a context where there is school choice, particularly, the accountability pressure introduced by the SEP law in Chile. This law offers extra funding for schools enrolling economically disadvantaged students, and the management of that extra funding, and the support for the development of improvement plans is

conditional on school's performance. Under this law, schools are classified in different performance categories according to a combination of rules that refer to test scores and other school quality indicators. Schools need to comply with certain thresholds in each of these rules to be classified as a low-, mid- or high-performing. This fact allowed me to use a regression discontinuity in a context where there are multiple rating of schools to assess the effect of receiving one classification versus the others. I used the binding score as instrumental variable to help me tackle the complexity of the formulas used to classify schools.

I fail to find systematic evidence that the accountability pressure on schools affects high-stakes outcomes or low-stakes outcomes. Furthermore, there is also not enough evidence to suggest there was any impact on school behavior, student and teacher mobility across schools. The lack of impact on the lowest performing schools is not consistent with the accumulated evidence of other studies that find a positive impact on high-stakes outcomes (Figlio & Rouse, 2006; Rockoff & Turner, 2008; Chiang, 2009; Allen & Burgess, 2012; Deming et al., 2013; Hussain, 2015). The lack of impact of better school classifications is not surprising. The accumulated evidence of other studies also fails to find an impact in those schools that just missed the best school classifications in New York City (Rockoff & Turner, 2008), Rhode Island (Weiner, Donaldson & Dougherty, 2016), or from another incentive policy in Chile (Rau & Contreras, 2009; Mizala & Urquiola, 2013), where schools receiving good accountability grades do not increase high-stakes test scores, and parents are also not responsive to the information on school quality.

In contrast to other research assessing the impact of the SEP law where its impact is assessed as a black box, this dissertation focused on the assessment of one of the components of the law,

that of school accountability. Particularly, it contributes to assessment of the impact of the pressure schools face as they are classified in different achievement categories.

Limitations

There are some caveats about these results. First, the conclusions of this study are limited to specific samples of schools. The estimated effects are local effects valid for schools that are close to the thresholds, and are valid for schools that comply with the rule of the ICE score. Another limitation of this study is that I cannot disentangle the effect of schools receiving certain classification (reputation), having the schools done themselves or with help the PME of the school, and the schools receiving more or less autonomy to manage the extra resources they receive per disadvantaged student. The only way these effects could be separated is if there were schools receiving different aspects of the treatment. For example, if there were schools receiving the classification of "autonomous", but where not given the autonomy to manage their resources or the Ministry gave them help to construct the PME. Or if there were schools classified as "in recovery" that were given autonomy to allocate the resources. Also the effect could be separated if schools classified as "emergent" were given autonomy to fully manage their resources or where not helped to construct the PME. By regulation, none of these situations should be happening. It is possible that the effect of these different components of the treatment may have cancelled each other out. It could be that schools positively benefit from the reputation of receiving a bad classification, but schools receive a negative impact of having less flexibility to allocate their resources. Another limitation of this study is that the time span between the year in which the schools receive the classifications and the year of the outcomes may not be enough to perceive any attempt of the schools to improve.

Another limitation of the study is that the SEP law had a clear mechanism through which it wanted to improve schools' performance, i.e., improving curriculum management, school leadership, school climate and educational resource management. Unfortunately the data available does not allow me to assess whether the policy in fact impacted any of these processes, nor if any impact on those processes may have affected students' test scores. I tested some potential mechanisms such as school responses, mobility of students and teacher employment. Nevertheless, the administrative data used for this purpose is very crude and may not capture improvement strategies used by the schools such as more training, or better curriculum.

These findings should be interpreted carefully. The lack of evidence of impact does not mean that the accountability component of the SEP does not impact schools' quality. This study does not assess the systemic effect of school accountability on the Chilean system, but the impact of consequences associated with the classification schools get. Therefore, this study does not disentangle the school accountability effect from the effect of extra resources found by other researchers about the impact of the introduction of the SEP law (Villarroel, 2012; MINEDUC, 2012; Correa et al., 2013; Mizala & Torche, 2013; Neilson, 2013; Navarro-Palau, 2015). Another precaution that should be taken when interpreting these results is that the impact of accountability pressure has been assessed four years after the law was first implemented. It is possible that accountability pressure had some impact the first years of the implementation and not years after.

Discussion of findings

There are several possible explanations of why there is no evidence of impact of accountability pressure. One possible but unlikely explanation is that all schools close to the thresholds felt the pressure of accountability –both that scored right above and right below,

pushing them all to improve. The other possibility is that the opposite happened: that no school felt any pressure to improve. My sense is that this is what happened, and I can think of several explanations. There could be problems with the assumptions behind school accountability, in its design, and/or implementation.

In terms of assumptions, it could be that schools did not work towards improving schools' academic outcomes because some of the basic assumptions discussed earlier in Chapter 2 did not hold. The fact that the classification formulas are so complex could make it hard for school personnel to correctly interpret the information about their students' achievement (*assumption 4*). It is possible the incentive offered was not desirable enough for school personnel, either in terms of form or magnitude or both (*assumption 5*), as the incentive targets "the school" as an organization and not individuals. It is also possible that the school had a weak leadership with a lack of knowledge on what to do and/or school personnel did not know alternative actions to improve school quality (*assumption 6*); or if they know such actions, its implementation did not lead to higher levels of performance (*assumption 7*). If with the accountability component of the SEP law the government wants to improve the quality of the schools, perhaps the government should look into the interventions implemented in Chile that have shown a positive impact on schools' performance (particularly in low-performing schools). For example, in the 1990s Chile introduced the P900 program. This program supplied low-performing schools with educational material and training for teachers, with no cash being transferred to schools. The evidence shows that this program had a positive and significant impact on students' outcomes (Tokman, 2002; Chay, McEwan & Urquiola, 2005). It may be worth assessing whether it may be more effective to expand or strengthen this type of program.

In terms of design, it is possible the incentives were not strong enough, or not clear enough, not driving schools to better allocate resources into what the Ministry thought were priority areas. It is clear the Ministry expected more resources to be allocated into the areas of curriculum management, leadership, school climate and resource management. And all extra resources had to be reported as being allocated in those areas. However, there are several difficulties when tracing the use of funds and their efficacy. First, schools were not restricted to substitute other funds allocated into the educational areas the SEP law aimed to improve. Therefore, it is possible that the resulting outcome is that schools did not allocate more resources into those areas, but simply substituted the funding source[75]. Furthermore, the fact that schools cannot charge extra fees to priority students, could have even decreased the expenditure on the priority areas or other educational areas. Second, the use of resources has been a matter of controversy, as almost 38% of resources incoming to schools are not properly accounted for or they are simply missing (Contraloria General de la Republica, 2014). Third, both government and schools face costs in tracing the money allocated into schools through the SEP law. The government has to monitor how the money is spent. The schools have to report and document how they are spending the money. It is worth thinking whether all these costs are worth the benefits schools are getting out the extra resources. Perhaps it would be cheaper and more efficient not to monitor the inputs the schools are getting, but the goals they are expected to achieve.

[75] Tsang and Levin (1983) discuss three major responses made by local governments to utilize intergovernmental grants. One response is that the local government uses the grants for the intended purposes. A second response is that local governments use the grant to substitute for local funding that would have been provided to support other educational services. A third response is that the grant is used to reduce local taxes or tuition fees to parents. Certainly, all three responses could have happened in the context of the SEP law. Unfortunately, I have no access to data in Chile about the allocation of resources to test these mechanisms.

In terms of implementation, there are at least four reasons why the pressure for a school classification may have debilitated the strength of incentives. First, sanctions for low-performers and low-performing classifications were postponed. One of the main incentives the law provided to low-performing schools (i.e., losing official recognition after four years of poor performance) was postponed by 12 years with the creation of the Quality Agency of Education. The first classifications of low-performing schools were not done two years after the implementation of the law as it was prescribed, but was delayed for four years, for no apparent reason. Such enforcement challenges, may compromise the credibility of the accountability system (Mintrop & Sunderman, 2009). If the government really wants to make schools accountable, then the government itself must do what is promised. If the government itself does not follow the rules they have created, how would it be expected for schools to follow those rules. One reason why the government seems to have postponed the punishment for low performing schools is because they were not clear on what would have happened with the students attending less-endowed schools. But clearly that is something the government should have thought of before even creating the law. This should serve as a lesson for the designers of the Quality Agency of Education and the Educational Superintendence.

Second, schools in all classifications were required to construct improvement plans (PME), but no clear consequences were attached to the compliance or not of the goals specified in such plans. The implementations of the plans were monitored, but not in accordance with the achievement of the goals described on the plan. What was monitored was whether the school was doing the "actions" committed to achieve those goals. This drives the focus away from the performance of the schools towards the input/processes of the school, which is just the opposite purpose of a performance-based program (Bush, Hough & Kirst, 2017). If the government had

really wanted schools to focus on the academic goals, then those are the ones that need to be monitored. Although just monitoring SIMCE test scores may be a little short sighted, they could monitor the attainment of other educational indicators in agreement with schools.

Third, the complexities of the formulas used to classify schools, and the weak communication of such formulas to the schools seems to undermine the capacity of schools to focus on such criteria. This could be seen as a positive or negative thing. If schools do not know the formulas in detail, then they may be less able to try to game the system. However, if they do not understand why they are classified the way they are, then how could they focus on improving those things. Furthermore, it does not seem necessary to make the rules so complex when there is clearly one rule that is binding for each type of classification. If the government wants schools to focus on improving specific indicators, then they may want to think in simplifying the indicators and explain them in simple terms.

Fourth, the information about the classification the school received was mainly communicated to school owners ("*sostenedores*"). School owners had to provide such information to parents, but it is hard to know whether they did or not. Although parents can have access to the school classifications through MINEDUC's website, it is hard to find, and it is not clear what it means (what an "emergent" school is, is not explained in the website where parents can get the data). Perhaps parents would have responded to information if it had been widely available, clear and informed by the Ministry directly and not just posted in a hard-to-find section of a website. Although there is evidence parents may not respond to information about high performing schools in Chile (Mizala & Urquiola, 2013). If the government had actually wanted to put pressure on schools to improve their quality by labelling the schools in different classifications they might have wanted to make information widely, easily and meaningfully

available to parents. The government may want to explore some strategies to inform schools and parents by exploring the accountability policy in Florida or California (Bush, Hough & Kirst, 2017).

Future research

The results suggest at least four avenues for future research. First, it would be interesting to disentangle the effect of reputation, support to create improvement plan and flexibility to allocate resources. It may be informative for the accountability scheme the Quality Agency of Education is building. Second, are the benefits of classifying schools worth the costs it carries for schools? Third, has the SEP law affected the intended school processes to improve? If so, has improvement in curriculum management or school leadership impacted students' achievement? Four, how has the SEP law impacted educational expenditures, overall and per school classification?

Considering how many countries are adopting school accountability programs, it is important for policymakers and practitioners to weigh the evidence into policy deliberations and program design. This dissertation suggests that it is hard for an accountability policy to have impact when the incentives are not precise, or when the actions prompted by the policy are not consistent throughout its design and implementation.

Tables

Table 1. Impact of school accountability on high-stakes outcomes.

Study	Independent variable	Outcome measure	Data (Place/Year/ Grade/Level)	Method	Findings
Richards & Sheu, 1992	Implementation of accountability system	Achievement gain (based on Readiness, BSAP and CTBS tests)	South Carolina, USA /1986-1988/ 1st to 11th grade/School-level	Analysis of trends years after introduction of policy	Modest improvements in student achievement (1% - 1.08% improvement), with large differences between schools' SES (lowest SES 2.45% and 1% improvement, whereas high SES 0.18% and 0.31%)
Klein, Hamilton, McCaffrey, & Stecher, 2000	Introduction of school accountability	TAAS math & reading	Texas, USA/1994-1998/4th & 8th grade/Student-level	Pre-post comparison	Increasing results in 4th grade (effect sizes ranging from 0.31 to 0.49) and 8th grade (ranging from 0.28 to 0.45).
Jacob, 2005	Introduction of accountability system	ITBS math & reading	Chicago, USA/1990-2000/3th, 6th & 8th grades/Student-level	Interrupted time series	Math and reading achievement increased after the introduction of the policy in comparison to prior trends. Reading increased somewhere between 0.026 and 0.24 sd, and math somewhere between -0.081 and 0.485.
			Chicago, USA/1993-2000/District-level	Difference-in-differences	Math and reading achievement increased after the introduction of the policy in comparison to other large urban districts. Math increased by 0.33 sd and reading 0.24 sd.
Figlio & Rouse, 2006	Accountability pressure to low performing schools	FCAT-SSS reading & math	Florida, USA / 1998-2000/3th, 4th & 5th grades / Student-level	Difference-in-differences	There is a small but significant increase in reading achievement (0.04), and a larger increase in mathematics (0.24).
Rockoff & Turner, 2008	Accountability pressure	Some NYC standardized test[76].	NYC, USA / 2006-2007/Elementary and Middle schools/School-level	Sharp regression discontinuity	There are significantly higher test scores for F (0.122[77] relative to a D grade) and D (0.122 sd relative to a C grade) schools in math and F schools in English (0.085 in reference to D schools).
Chiang, 2009	Imposition of school sanction threats	FCAT	Florida, USA/ 2002-2003/3th, 4th , 5th and 6th grade/Student-level	Sharp regression discontinuity	There is a significant impact on 4th grade results in math and reading (0.118 sd and 0.122 sd respectively). The results do not seem to remain in 5th and 6th grade (only the impact in math in 6th grade remained 0.109 sd)

[76] Test not specified on paper.

[77] Effect size calculated taking the difference from the effect of F-D (to have D as a reference) and then dividing by the standard deviation of the achievement measure (17.2 in math, and 21.2 in reading).

Table 1, Continued.

Study	Independent variable	Outcome measure	Data (Place/Year/ Grade/Level)	Method	Findings
Neal & Schanzenbach, 2010	Introduction of accountability system	ISAT math & reading	Chicago, USA/2001-2002/5th grade/Student-level	Difference-in-differences	Increase reading and math scores among the students in the middle of the achievement distribution and not among the students at the left tail of the achievement distribution (from 3th to 9th decile increases between 0.053 and 0.134 in math, and from 3th to 9th decile increases between 0.038 and 0.09 in reading[78]).
		ITBS math & reading	Chicago, USA/1996-1998/5th grade/Student-level		Increase reading and math scores among the students in the middle of the achievement distribution and not among the students at the left tail of the achievement distribution (from second to 9th decile increases between 0.003 and 0.006 in mean test scores in math, and from 4th to 8th decile increases between 0.004 and 0.005 in reading[79]). There is even a negative impact on students from the first decile of the achievement distribution (-0.11 sd).
Allen and Burgess, 2012	Accountability pressure on schools failing inspection	GSCE Math & English test	England/2002-2011/Students age 16	Fuzzy regression discontinuity	Just-failing schools improve scores over the following two to three years. The effect size is moderate at around 0.1 standard deviations. The impact is mainly on middle and top end of the ability distribution.
				Regression discontinuity and difference-in-differences	Just-failing schools improve scores over the following two to three years. The effect sizes are small ranging between 0.03 and 0.06standard deviations.
Deming, Cohodes, Jennings and Jencks, 2013	Accountability pressure	Passed high-stakes test on time (%)	Texas, USA/1994-2010/8th grade students followed until 25 years old/Student-level	School fixed effects	High schools that face the possibility of being rated as low performing schools increase the probability that students pass high stakes exams on time (0.7%). The impact is greater on low achieving students (students that failed the year previous to the test). They increase the passing rates of the high-stakes test by 1.5%.

[78] Effect size calculated dividing the reported difference in mean scores by the standard deviation of the ISAT (15 sd).

[79] Standard deviation of the test not reported in paper. Effect size calculated dividing the reported mean difference by 21.06, the standard deviation of ITBS according to Rouse & Figlio (2006).

Hussain, 2015	Accountability pressure on schools failing inspection	Key Stage 2 Math & English	England/2006-2009/Age 11/Student level	Difference-in-differences	Students from failing schools improve their test scores around 0.1 standard deviations. The largest gains are for students in the bottom quartile of the age-7 test scores distribution with gains climbing up to a 0.2 sd.

Table 1, Continued.

Study	Independent variable	Outcome measure	Data (Place/Year/ Grade/Level)	Method	Findings
Weiner, Donaldson and Dougherty, 2016	Accountability pressure to schools that just missed being classified as high-performers	NECAP math & reading	Rhode Island, USA/2010-2014/3th to 8th grade/Grade level	Fuzzy regression discontinuity	There is no evidence just missing high-performance status improves students' performance. However, when the level of competence with other high-performing schools is assessed, then the study finds that schools that just missed being classified as high-performing and are in a context with several other high performing schools, then there is an improvement in the school's outcomes. This does not happen in schools in a setting where there are not many high-performing schools.

Source: Author.

Table 2. Impact of school accountability on low-stakes outcomes.

Study	Independent variable	Outcome measure	Data (Place/Year/ Grade/Level)	Method	Findings
Klein, et al., 2000	Introduction of accountability system	NAEP math & reading	Texas, USA/1994-1998/4th & 8th grade/Student-level	Pre-post comparison	Small increase in 4th grade scores (0.13 to 0.15). No increase in 8th grade scores.
Jacob, 2005	Introduction of accountability system	IGAP math	Chicago, USA/1990-2000/3th, 6th & 8th grades/Student-level	Interrupted time series	Decline in low stakes scores among third graders of roughly 0.13 standard deviations, had no effect on scores among sixth graders, and increased eight grade scores by roughly 0.26 standard deviations.
		IGAP math &reading	Chicago, USA/1993-2000/District-level	Difference-in-differences	No effect on low stakes performance among 3th and 6th graders, but did increase the low stakes scores in 8th grade for reading (0.211 standard deviations) as well as for math (0.164 sd).
Figlio & Rouse, 2006	Introduction of accountability pressure to low performing schools	NRT reading & math	Florida, USA / 1998-2000/4th & 5th grades / Student-level	Difference-in-differences	Impact in low stakes tests is positive and statistically significant (0.106 in math 5th grade and 0.03 in reading 4th grade[80]), but considerably smaller than those found on the high stakes tests.
Chiang, 2009	Imposition of school sanction threats	Stanford test math & reading	Florida, USA/ 2002-2003/3th & 4th grade/Student-level	Sharp regression discontinuity	Impact of sanction threats on low stakes tests are negligible for both math and reading.
Deming, Cohodes, Jennings and Jencks, 2013	Accountability pressure	Graduated high school (%), Total math credits in high school (#)	Texas, USA/1994-2010/8th grade students followed until 25 years old/Student-level	School fixed effects	High schools that face the possibility of being rated as low performing schools increase the percentage of high school graduates (0.9%) and increase total math credits taken in high schools (0.06) . High-performing schools had a negative impact on high school graduation rates (0.9%). The impact is greater on low achieving students (students that failed the year previous to the test). They increase the percentage of high school graduates (1%) and increase the number of credits taken in math in high school (0.073). Schools that will probably be ranked as recognized may even show negative effects on students who failed the exam the previous year (probably because they re-categorized these students as having special needs).

Source: Author.

[80] Effect size calculated by dividing the coefficient estimate by the test standard deviation of 21.06

Table 3 Impact of school accountability on long-run outcomes.

Study	Independent variable	Outcome measure	Data (Place/Year/ Grade/Level)	Method	Findings
Deming, Cohodes, Jennings and Jencks, 2013	Accountability pressure	Passed high-stakes test on time (%), Graduated high school (%), Total math credits in high school (#), Attended college (%), Attended 4 year college (%), BA degree (%), Earnings at 25 ($), Earnings between 23-25 ($)	Texas, USA/1994-2010/8th grade students followed until 25 years old/Student-level	School fixed effects	Students in school cohorts at risk of being graded low performing were about 1% more likely to attend college than high performing cohorts and 12% more likely to enroll in 4 year colleges. Students in school cohorts at risk of being graded low performing earned between 23 and 25 years old about $459 USD more than high performing cohorts. Increases in postsecondary attainment are much larger for lower scoring students (those who have failed in 8th grade math tests) in low performing cohorts. Low achieving students in low performing schools attended 4 year college about 14% more than low performing students in an acceptable performance school. Low achieving students in low performing schools earned about $518 USD more than low achieving students in schools with acceptable performance. Low achieving students in high performing schools earned about $1200 USD less than low achieving students in schools with acceptable performance.

Source: Author

Table 4. Impact of school accountability on school behavioral response.

Study	Independent variable	Outcome measures	Data (Place/Year/ Grade/Level)	Method	Findings
Jacob, 2005	Introduction of accountability system	Number of students in special education, Number of scores reported, Number of students taking the test, Retention rates in grades prior to high-stakes grades	Chicago, USA/1990-2000/3th, 6th & 8th grades/Student-level	Interrupted time series	The introduction of the policy increased proportion of students in special education (1%), decreased number of students tested (0.6%), increased the grade retention among students in 1st and 2nd grade by 2.3% (64% more than the baseline 3.6%), and by 1.7 in 4th, 5th and 7th grade (130% more than the baseline 1.3%). There is no evidence the introduction of the policy affected the number of scores reported.
Cullen & Reback, 2006	Pressure to account for certain subgroups of student	Exemption rates	Texas, USA/1993-1998/3th, 4th, 8th and 10th grades/Student-level	Difference-in-differences	Campuses target exemptions toward student subgroups when they are likely to prevent the campus from earning a higher rating. Exemptions increase between a 0.6 and 1.7 percent. The exemptions can be larger in Hispanic and Black subgroups (between 1.2 to 2.1 percent, and between 1.3 to 3.7 percent respectively).
Jacob, 2007	Tests gap by year	Item response (correct or wrong) on NAEP (low stakes test) and TAAS (high stakes test)	Texas/1996-2000/4th and 8th grades/Item level	Comparison of performance trends	Differential improvement in state test cannot be explained by changes in demographics of test-takers, or format of the tests (open response vs. multiple choice items; use of calculators; timing of the test). Different skills assessed may explain the differences in 4th graders, but not in 8th graders.
Rockoff & Turner, 2008	Accountability pressure	Percentage of students tested in math and English	NYC, USA / 2006-2007/Elementary and Middle schools/School-level	Sharp regression discontinuity	There is no evidence between the accountability grade the school receives and the percentage of students tested in math and English.
Chiang, 2009	Imposition of school sanction threats	School expenditures (Total school costs per pupil, Total instructional costs pp, Total non-instructional costs pp, Share of costs devoted to instruction, Ratio of FTE students to FTE staff units), Instructional costs (Teacher salaries and benefits pp, Contracted service costs pp, Instructional materials costs pp, Instructional equipment costs pp), Non-instructional costs (Pupil support costs pp, Media center costs pp, Instructional and curricular development costs pp, Teacher training costs pp, School administration costs pp, Plant operation costs pp, Plant maintenance costs pp)	Florida, USA/ 2002-2003/3th & 4th grade/Student-level	Sharp regression discontinuity	Schools increase expenditure on instructional equipment costs ($151 USD per pupil) and instructional and curricular development costs ($189 USD per pupil). There is no evidence of changes in any other type of expenditure.

Table 4, Continued.

Study	Independent variable	Outcome measures	Data (Place/Year/ Grade/Level)	Method	Findings
Craig, Imberman & Purdue, 2013	Face rating reduction of the school due to changes in the accountability tests	Total expenditures per student, Resources per student: categorical expenditures (instruction, leadership/curriculum/staff development, counseling and social work services, extracurricular activities, student-faculty ratios)	Texas, USA/2002-2006/School-level	Rating shock	The authors find that school districts increase instructional budgets (increase total expenditures per student by $73USD, increase in instructional budget by $66 USD, student-teacher ratios decreased by 0.3%) for teachers if there was an increase likelihood of a lower accountability grade (not in the lowest level, but in the second lowest level of rating. This increase is no longer found 3 years after the shock.
Rouse , Hannaway, Goldhaber & Figlio, 2013	Accountability pressure to low performing schools	Policy changes in several domains to: improve low performing students, lengthen the instructional time, reduce class size for math, reading and writing, narrow the curriculum, change scheduling system, improve low performing teachers, improve teacher resources, improve teacher incentives, change in level of control for teachers, district and principal, and others.	Florida, USA/1999-2004/School-level	Sharp regression discontinuity	Schools that face pressure focus on low-performing students, increase the amount of time devoted to instruction organize learning differently, increase resources available to teachers, and decrease principal control.
Hussain, 2015	Accountability pressure on schools failing inspection	Number of low-ability students in the test-taking pool	England/2006-2009/Age 11/Student level	Difference-in-differences	There is no evidence failing schools exclude low-ability students

Source: Author.

Table 5. Impact of school accountability on student body composition.

Study	Independent variable	Outcome measures	Data (Place/Year/ Grade/Level)	Method	Findings
Hart & Figlio, 2015	Introduction of school accountability policy	Average years of completed education of mothers, Average maternal age, Share of kindergarten class with married parents at birth, share of class that are low income (need subsidized lunch)	Florida, USA/1997-2002/Kindergarten/Student-level	Difference-in-differences	High SES parents were particularly responsive to the introduction of schools grades (greater increases in the average maternal education of schools post reform for A and B schools, and less for D and F schools; similar patterns found in maternal age; but no evidence was found in whether the mother was married at time of birth, or whether the child received free or reduced lunch). High performing schools had an increase in the SES of the families among the Kindergarten students. The effect is stronger when there are nearby alternatives for the school (when there is actually choice) and where nearby alternatives are poorer performing schools.

Source: Author.

Table 6. Impact of school accountability on teacher body composition.

Study	Independent variable	Outcome measures	Data (Place/Year/ Grade/Level)	Method	Findings
Clotfelter, Ladd, Vigdor & Aliaga Diaz, 2004	Introduction of accountability pressure to low performing schools	Teacher retention rates, Hiring rates, Proportion of low quality teachers (measured as teachers with no experience, and as teachers ho graduated from uncompetitive colleges)	North Carolina, USA/Elementary schools/Teacher –level	Difference-in-differences	Labeling of schools as low-performing increases the probability of departure for an experienced teacher by about 25%. For a new teacher, the changes in the probabilities of departure are around the same, although the baseline level of departure was higher than for experienced teachers. There is no evidence that the introduction of accountability has decreased the quality of the teachers in the low-performing schools.
Feng, Figlio & Sass, 2010	Whether the school was upward or downward shocked after a change in grading formula of schools	Likelihood that a teacher leaves his or her school before the end of the year after the change of school grade	Florida, USA/1995-2003/elementary, middle and high schools/Teacher-level	Difference-in-differences	The effect of schools being classified lower than expected affects teacher mobility increasing it by 11-12% in comparison to those schools that were not shocked. The effect of schools being classified higher than expected does not significantly affect teacher mobility, although the numbers indicate it decreases slightly.

Source: Author.

Table 7. Construction of the Education Quality Index (ICE).

			Private voucher school	Public school
SIMCE average		Average of three last scores of SIMCE in 4th grade (math, language and sciences)	70%	
Other quality indicators			30%	
	Approval rates	Proportion of students approved.	25%	25%
	Retention rates	Proportion of students enrolled until the end of the academic year.	25%	25%
	SNED Improvement	Improvement on working conditions and functioning of the school. Count of sanctions to the school weighted by how severe they are.	20%	17%
	SNED Initiative	Schools' activities and initiatives and commit external actors in its educational practices. Data from survey filled by 'school principal'.	15%	13%
	SNED Integration	Participation of teachers and parents in the construction of the educational project of the school. Data from survey filled by 'school principal' and parental questionnaire that accompanies the SIMCE test.	15%	13%
	Teacher evaluation	Teacher evaluation of public school teachers. Index: (Number of excellent teachers + Number of competent teachers – Number of incompetent teachers) / (Total number of teachers assessed). Data from system of teacher evaluation.	-	7%

Note: Data from the decree 293, Technical Report of *Proceso de Clasificacion SEP*, Technical Report of *SNED 2012-2013.*

Table 8. Adoption of SEP law through the years.

Schools		Years							
		2008	2009	2010	2011	2012	2013	2014	2015
Public	#	4964	5013	4937	4823	4801	5058	5001	4960
	% from total	96.75	96.14	99.00	94.9	87.35	93.46	94.13	94.42
	Total public	5131	5214	4987	5082	5496	5412	5313	5253
Private Voucher	#	1635	2063	2182	2455	2658	2909	2950	3008
	% from total	42.30	45.4	54.6	50.86	44.85	48.52	48.82	49.69
	Total PV	3865	4544	3996	4827	5927	5995	6042	6053

Note: Data from SEP database of Centro de Estudios MINEDUC (2008-2015). Total schools are those that have at least 1 priority student and have elementary education.

Table 9. Number of priority students throughout the years.

Schools		Years							
		2008	2009	2010	2011	2012	2013	2014	2015
Public	# of priority students	258886	425956	447950	483560	769331	889695	839350	754055
	% of priority students from total enrollment	35.58	51.31	58.91	57.71	66.46	76.34	73.56	67.43
Private Voucher	# of priority students	128044	323530	361556	421746	768811	935204	922939	859640
	% of priority students from total enrollment	17.24	28.78	36.06	32.17	48.48	57.51	56.86	52.05

Note: Data from SEP database of Centro de Estudios MINEDUC (2008-2015). Total schools are those that have at least 1 priority student, and have elementary education. The number of priority students and the percentage of priority students per school is calculated regardless of whether the school adopted or not the SEP law. Note that throughout the years the number of priority students is rolling upwards because each extra year includes new grades into the SEP law.

Table 10. Schools classified or not by formula.

	Years			
	2012	2013	2014	2015
Uses formula (%)	2901 (38.89)	2867 (35.99)	2859 (36.00)	2888 (36.25)
Less than 2 SIMCE measures	3012	3528	3195	2335
20 students or less take the test	1546	1568	1880	2741
Do not present PME	.	4	7	2
Total	7459	7967	7941	7966

Note: Data from SEP classification (2012-2015). Data about schools that do not present PME for 2012 is not available. "Uses formula" means the schools use the formula to get the classification they have. "Less than 2 SIMCE measures" means that the school has less than 2 years of valid SIMCE scores. This includes all schools that do not have 4th grade, and all new schools. "Less than 20 students take the test" means that during the 3 years for which SIMCE scores are considered and for the three different tests, on average the schools have no more than 20 students taking the tests. This includes all schools that are multiple-grade schools, geographically isolated schools and schools with three teachers or less. "Do not present PME" means the school was once classified using the formula, but as the school did not present the PME on time, the school was re-classified as "In recovery", regardless of the academic results of the school.

Table 11. School classifications through the years.

	Years			
Classification	2012	2013	2014	2015
Autonomous	1069	1071	1142	1155
Emergent	1644	1621	1663	1662
In recovery	188	175	54	71
Total	2901	2867	2859	2888

Note: Data from SEP classification (2012-2015). All SEP schools that have elementary education that are classified using the formula.

Table 12. School classification changes.

	Classification on year t+1 2013			Classification on year t+1 2014			Classification on year t+1 2015		
Classification on year t	Auto.	Emerg.	In Rec.	Auto.	Emerg.	In Rec.	Auto.	Emerg.	In Rec.
Autonomous	809	243		848	209		895	241	
Emergent	210	1350	62	241	1358	7	202	1423	31
In recovery		82	103	1	123	45		21	33

Note: Data from SEP classification (2012-2015). All SEP schools that have elementary education that are classified using the formula.

Table 13. Impact of SEP law on high-stakes outcomes.

Study	Independent variable	Outcome measure	Data (Place/Year/ Grade/Level)	Method	Findings
Villarroel, 2012	Introduction of SEP law	SIMCE math and language	Chile/2007-2010/4th grade/School-level	Propensity score matching and Difference-in-differences	The law increases language and math scores. The effect sizes range between 0.11 and 0.18 in math and 0.07 and 0.11 in language. The impact is larger for schools with more proportion of priority students and for schools that have implemented improvement plans for longer time.
MINEDUC, 2012	Introduction of SEP law	SIMCE math and language	Chile/2006-2011/4th grade/School-level	Difference-in-differences	After 4 years of implementation the SEP law has had a positive and significant effect on math and language in private voucher schools . With effect sizes ranging from 0.08 to 0.11 in and 0.05 to 0.07 respectively[81]. Each extra year participating in the SEP law increases the achievement by 0.03 in math and 0.02 in language.
Correa, Inostroza, Parro, Reyes & Ugarte, 2013	Introduction of SEP law	SIMCE math and language	Chile/2006-2011/4th grade/School-level	Difference-in-differences	After 4 years of implementation the SEP law has had a positive and significant effect on math and language in private voucher schools. With effect sizes of 0.08 and 0.05 respectively[34]. Each extra year participating in the SEP law increases the achievement by 0.05 in math and 0.02 in language.
Mizala & Torche, 2013	Introduction of SEP law	SIMCE math and language	Chile/2006-2011/4th grade/School-level	School fixed effects	Private voucher schools increase their average test scores once they adhere to the SEP Law. The effect sizes in math range from 0.08 to 0.1 and from 0.07 to 0.08 in language[34]. Each extra year of participation in the SEP law impacted positively on the outcomes. The effect of the SEP law is heterogeneous for schools with different average socioeconomic levels of the families. Schools that enroll students from lower SES have a greater impact of the law than other schools. In higher SES schools the impact of the law is negligible.
Neilson, 2013	Introduction of SEP law	Average SIMCE of math and language	Chile/2004-2011/4th grade/Student-level	Difference-in-differences	The policy impact of the targeted voucher program was to increase test scores of the poorest 40% of students by 0.2 standard deviations after the fourth year of its implementation.
Navarro-Palau, 2015	Increased school choice	Average SIMCE of math and language	Chile/2005-2014/4th grade/Student-level	Regression discontinuity and Difference-in-differences	There are no effects on average test scores for students with mother that completed high school (which show higher tendency to switch schools to a private voucher school due to the introduction of the SEP law). In contrast, the average test scores of students with low educated mothers (students that mostly stay in public schools), increased (0.08 sd).

[81] Effect sizes calculated dividing the reported coefficients by the standard deviation of SIMCE (50).

Source: Author.

Table 14. Determinants of "autonomous" classification in 2012.

	(1) SIMCE t-2	(2) SIMCE t-3	(3) SIMCE t-4	(4) P250 t-2	(5) P250 t-3	(6) P250 t-4	(7) P300 t-2	(8) P300 t-3	(9) P300 t-4	(10) ICE t-1
Above threshold	0.30***	0.33***	0.28***	0.27***	0.34***	0.30***	0.29***	0.29***	0.30***	0.53***
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Running variable	0.01***	0.01***	0.01***	1.26***	1.05***	1.24***	1.88***	1.89***	1.94***	0.24***
	(0.00)	(0.00)	(0.00)	(0.08)	(0.07)	(0.08)	(0.12)	(0.11)	(0.12)	(0.01)
Observations	2888	2799	2867	2888	2798	2867	2888	2799	2867	2901
R^2	0.445	0.479	0.450	0.434	0.461	0.455	0.422	0.438	0.415	0.677
F	463.07	513.78	467.71	442.72	477.09	478.62	420.52	434.90	406.61	1211.30

***Note*:** The models are all linear probability models of whether the school is classified as "autonomous" or not on year 2012 (equation 5 without the interaction term). The independent variables "Above threshold" are dummy variables that indicate whether the school scores above (1) or below (0) the median of the SES group of reference in the rule mentioned at the top of the column. All models control for the type of school, SES level of the school, whether the school is urban, and for the enrollment rates in 4th grade. Each cell contains the coefficients and standard error on parenthesis. The schools included are those SEP schools that are classified according to the classification formula for the full sample.
* p<0.1, **p<0.05, ***p<0.01

Table 15. Determinants of school classification as "in recovery" in 2012 & 2013 rounds pooled.

	(1) SIMCE t-2	(2) SIMCE t-3	(3) SIMCE t-4	(4) P250 t-2	(5) P250 t-3	(6) P250 t-4	(7) ICE t-1
Below threshold	0.39***	0.33***	0.27***	0.51***	0.41***	0.42***	0.97***
	(0.01)	(0.01)	(0.001)	(0.02)	(0.02)	(0.01)	(0.00)
Running variable	-0.00***	-0.00***	-0.00***	-0.35***	-0.26***	-0.26***	-0.01***
	(0.00)	(0.00)	(0.00)	(0.02)	(0.02)	(0.02)	(0.00)
Observations	5722	5647	5611	5732	5646	5611	5763
F	309.47	279.72	253.48	276.27	244.66	280.82	9348.24

***Note*:** The models are all linear probability models of whether the school is classified as "in recovery" or not on year 2012 and 2013 (equation 7 without the interaction term). The independent variables "Below threshold" are dummy variables that indicate whether the school scores above (0) or below (1) the threshold in the rule mentioned at the top of the column. All models control for the type of school, SES level of the school, whether the school is urban, and for the enrollment rates in 4th grade. Each model also contains a fixed effect per round. Each cell contains the coefficients and standard error on parenthesis. The schools included are those SEP schools that are classified according to the classification formula for the full sample in both years.

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

Table 16. Descriptive statistics of schools around the cutoff of the "autonomous" classification in 2012.

	(1)	(2)
	Below the cutoff	Above the cutoff
Number of schools	1149	1063
SEP classification		
Autonomous	0.000	0.704
Emergent	0.980	0.296
In recovery	0.020	0.000
Urbanicity		
Urban schools	0.932	0.877
Rural schools	0.068	0.123
Type of school		
Municipal	0.645	0.564
Private voucher	0.355	0.437
SEP participation		
Average years in SEP until 2012	3.676	3.673
Enrollment		
Average enrollment	533.831	583.232
% of enrollment in grade 4	0.095	0.096
% of enrollment in grade 8	0.104	0.100
Student characteristics		
Low SES (%)	0.094	0.116
Mid-low SES (%)	0.484	0.463
Mid SES (%)	0.350	0.355
Mid-High, High SES (%)	0.072	0.062
% of beneficiary students	0.496	0.502
% of priority students	0.543	0.545
% of students with special needs	0.066	0.050
School classification scores for 2012's classification		
SIMCE score t-2	242.37	257.32
% scoring above 250 on t-2	0.447	0.572
% scoring above 300 on t-2	0.129	0.195
SIMCE score t-3	238.33	255.32
% scoring above 250 on t-3	0.413	0.551
% scoring above 300 on t-3	0.117	0.189
SIMCE score t-4	235.68	251.19
% scoring above 250 on t-4	0.396	0.519
% scoring above 300 on t-4	0.108	0.168
ICE index on t-1	-0.221	0.371

Note: Descriptive statistics are presented for schools above and below the ICE index cutoff. The schools included are those SEP schools that are classified using the classification formula, and that are within the 0.706 bandwidth.

Table 17. Descriptive statistics of schools around the cutoff of the "in recovery" classification in 2012 & 2013 rounds pooled.

	(1) Below the cutoff	(2) Above the cutoff
Number of observations	276	787
SEP classification		
Autonomous	0.000	0.000
Emergent	0.004	0.977
In recovery	0.996	0.002
Urbanicity		
Urban schools	0.964	0.916
Rural schools	0.036	0.084
Type of school		
Municipal	0.649	0.710
Private voucher	0.351	0.290
SEP participation		
Average years in SEP until 2012	3.830	3.753
Enrollment		
Average enrollment	414.181	462.653
% of enrollment in grade 4	0.094	0.093
% of enrollment in grade 8	0.104	0.106
Student characteristics		
Low SES (%)	0.304	0.188
Mid-low SES (%)	0.623	0.658
Mid SES (%)	0.073	0.153
Mid-High, High SES (%)	0.000	0.001
% of beneficiary students	0.621	0.614
% of priority students	0.685	0.664
% of students with special needs	0.111	0.103
Average school classification scores for 2012 and 2013 rounds of classification		
SIMCE score t-2	218.38	230.30
% scoring above 250 on t-2	0.253	0.343
SIMCE score t-3	215.82	225.77
% scoring above 250 on t-3	0.241	0.312
SIMCE score t-4	211.57	222.84
% scoring above 250 on t-4	0.213	0.291
ICE index on t-1	-1.240	-0.774

Note: Descriptive statistics are presented for schools above and below the ICE index cutoff. The schools included are those SEP schools that are classified using the classification formula, and that are within the 0.467 bandwidth.

Table 18. Mean of high- and low- stake outcome variables above and below the ICE index threshold of the "autonomous" classification in 2012.

	(1) Below the cutoff	(2) Above the cutoff
Panel A: High stakes outcomes		
4th grade Math scores on year t	245.71	259.20
4th grade Math scores on year t+1	240.55	252.66
4th grade Language scores on year t	253.27	264.45
4th grade Language scores on year t+1	249.57	259.85
4th grade Science scores on year t	242.26	254.48
4th grade Science scores on year t+1	239.87	250.23
Panel B: Low stakes outcomes		
8th grade Math scores on year t+1	242.47	253.89
8th grade Language scores on year t+1	238.66	250.84
8th grade Science scores on year t+1	254.44	265.39

Note: Mean outcomes are presented for schools above and below the ICE index cutoff. The schools included are those SEP schools that are classified using the classification formula, and that are within the 0.706 bandwidth.

Table 19. Mean of high- and low- stake outcome variables above and below the ICE index threshold of the "in recovery" classification in 2012 & 2013 rounds pooled.

	(1) Below the cutoff	(2) Above the cutoff
Panel A: High stakes outcomes		
4th grade Math scores on year t	227.40	232.69
4th grade Math scores on year t+1#	228.88	232.07
4th grade Language scores on year t	235.86	241.55
4th grade Language scores on year t+1#	236.61	239.92
4th grade Science scores on year t	224.88	230.53
4th grade Science scores on year t+1#	228.27	231.37
Panel B: Low stakes outcomes		
8th grade Math scores on year t*	228.06	232.37
8th grade Math scores on year t+1#	228.80	232.95
8th grade Language scores on year t*	223.36	227.73
8th grade Language scores on year t+1#	224.36	228.65
8th grade Science scores on year t*	239.05	244.26
8th grade Science scores on year t+1#	240.01	244.02

Note: Mean outcomes are presented for schools above and below the ICE index cutoff. The schools included are those SEP schools that are classified using the classification formula, and that are within the 0.467 bandwidth.
#Outcome for round of 2012 only. *Outcome only available for round of 2013.

Table 20. Mean of school behavioral response, student and teacher body composition outcome variables above and below the ICE index threshold of the "autonomous" classification in 2012.

	(1) Below the cutoff	(2) Above the cutoff
Panel A: School behavioral response		
Tests taken on year t in 4th grade	131.80	146.15
Tests taken on year t+1 in 4th grade	126.32	141.87
Students with special needs on year t	34.71	28.03
Students with special needs on year t+1	53.42	52.15
Students retained in 3th grade on year t	1.91	1.87
Students retained in 3th grade on year t+1	1.70	1.622
Students approved in 3th grade on year t	44.67	50.30
Students approved in 3th grade on year t+1	43.68	49.79
Students switching schools in 3th grade on year t	3.39	2.87
Students switching schools in 3th grade on year t+1	3.32	2.81
Proportion of students scoring above 250 on year t	0.482	0.581
Proportion of students scoring above 250 on year t+1	0.449	0.541
Proportion of students scoring above 300 on year t	0.147	0.207
Proportion of students scoring above 300 on year t+1	0.117	0.165
ICE index on year t	-0.234	0.337
ICE index on year t+1	-0.072	0.342
Teacher evaluation t	0.679	0.754
Teacher evaluation t+1	0.681	0.750
SNED initiative on t	79.21	85.57
SNED initiative on t+1	79.29	85.62
SNED improvement on t	92.67	92.02
SNED improvement on t+1	92.65	91.97
SNED integration on t	69.07	78.04
SNED integration on t+1	69.09	78.04
Approval rates on t	0.940	0.952
Approval rates on t+1	0.944	0.953
Retention rates on t	0.979	0.987
Retention rates on t+1	0.978	0.987
Panel B: Student body composition		
Total school enrollment on year t	533.83	583.23
Total school enrollment on year t+1	525.18	582.15
School enrollment in 1st grade on year t	45.91	51.60
School enrollment in 1st grade on year t+1	46.20	53.09
School enrollment in 7th grade on year t	52.42	55.44
School enrollment in 7th grade on year t+1	52.82	57.15

Table 20, Continued.

	(1) Below the cutoff	(2) Above the cutoff
4th grade SES level on year t: Low SES (%)	0.094	0.116
4th grade SES level on year t+1: Low SES (%)	0.117	0.114
4th grade SES level on year t: Mid-low SES (%)	0.484	0.463
4th grade SES level on year t+1: Mid-low SES (%)	0.513	0.469
4th grade SES level on year t: Mid SES (%)	0.350	0.359
4th grade SES level on year t+1: Mid SES (%)	0.323	0.370
4th grade SES level on year t: Mid-High, High SES (%)	0.072	0.062
4th grade SES level on year t+1: Mid-High, High SES (%)	0.046	0.047
% of sch with enrollment fees on year t+1: free	0.90	0.89
% of sch with enrollment fees on year t+1: 1 to 10 thds pesos	0.09	0.09
% of sch with enrollment fees on year t+1: 10 to 25	0.005	0.009
% of sch with enrollment fees on year t+1: 25 to 50	0.004	0.002
% of sch with enrollment fees on year t+1: 50 to 100*	0.000	0.001
% of sch with monthly fees on year t+1: free	0.77	0.71
% of sch with monthly fees on year t+1: 1 to 10 thds pesos	0.06	0.07
% of sch with monthly fees on year t+1: 10 to 25	0.11	0.14
% of sch with monthly fees on year t+1: 25 to 50	0.04	0.07
% of sch with monthly fees on year t+1: 50 to 100**	0.001	0.002
Beneficiary students enrolled on year t	239.22	261.81
Beneficiary students enrolled on year t+1	289.26	323.14
Panel C: Teacher body composition		
Teachers teaching on year t	25.51	26.76
Teachers teaching on year t+1	26.33	28.08
Teaching hours on year t	838.14	893.66
Teaching hours on year t+1	872.07	946.82

Note: Mean outcomes are presented for schools above and below the ICE index cutoff. The schools included are those SEP schools that are classified using the classification formula, and that are within the 0.706 bandwidth. *Although schools could have reported enrollment fees greater than 100, there were no observations on this level on this sample. **Although schools could have reported monthly fees greater than 100, there were no observations on this level on this sample.

Table 21. Mean of school behavioral response, student and teacher body composition outcome variables above and below the ICE index threshold of the "in recovery" classification. 2012 & 2013 classification rounds pooled.

	(1) Below the cutoff	(2) Above the cutoff
Panel A: School behavioral response		
Tests taken on year t in 4th grade	95.66	108.27
Students with special needs on year t	44.47	44.72
Students retained in 3th grade on year t	2.14	1.94
Students approved in 3th grade on year t	32.50	37.70
Students switching schools in 3th grade on year t	4.13	3.62
Proportion of students scoring above 250 on year t	0.333	0.377
Proportion of students scoring above 300 on year t	0.072	0.091
ICE index on year t	-0.623	-0.984
Teacher evaluation t	0.584	0.646
SNED initiative on t	68.01	72.53
SNED improvement on t	91.77	91.61
SNED integration on t	55.34	61.45
Approval rates on t	0.917	0.928
Retention rates on t	0.951	0.968
Panel B: Student body composition		
Total school enrollment on year t	414.18	462.65
School enrollment in 1st grade on year t	35.29	40.44
School enrollment in 7th grade on year t	42.21	48.33
4th grade SES level on year t: Low SES (%)	0.304	0.188
4th grade SES level on year t: Mid-low SES (%)	0.623	0.658
4th grade SES level on year t: Mid SES (%)	0.073	0.153
4th grade SES level on year t: Mid-High, High SES (%)	0.000	0.001
Beneficiary students enrolled on year t	247.48	270.26
Panel C: Teacher body composition		
Teachers teaching on year t	22.33	24.08
Teaching hours on year t	727.86	792.44

Note: Mean outcomes are presented for schools above and below the ICE index cutoff. The schools included are those SEP schools that are classified using the classification formula, and that are within the 0.467 bandwidth.

Table 22. First stage models for schools being classified or not as "autonomous". Year 2012.

	(1) Baseline	(2) With controls
Above ICE $_{t-1}$	0.30***	0.29***
	(0.03)	(0.03)
ICE $_{t-1}$	-0.00**	-0.001
	(0.00)	(0.00)
Above ICE $_{t-1}$ * ICE $_{t-1}$	1.36***	1.35***
	(0.06)	(0.06)
Observations	2209	2209
F	2885.91	2528.86

Notes: The table shows resulting coefficients from equation 5 estimated with a linear probability model. The cells contain the main coefficients and the robust standard errors in parenthesis. The schools included are those SEP schools that are classified according to the classification formula, and that are within the 0.706 bandwidth. Covariates control for whether the school is in an urban area, the type of school, schools SES level, and the enrollment rates in 4th grade.
* p<0.1, **p<0.05, ***p<0.01

Table 23. Specification check of RD estimate of receiving a school classification of "autonomous" on mathematics outcomes in 4th grade on year t. Year 2012.

	(1)	(2)	(3)	(4)
Bandwidth:	[-1.2 ; 1.2]	[-.9 ; .9]	[-.706 ; .706]	[-.5 ; .5]
Polynomial of order:				
One				
Coeff. (SD)	2.77 (2.42)	2.72 (2.46)	0.81 (2.73)	1.10 (3.42)
F (p-value)	147.56 (0.000)	143.85 (0.000)	119.70 (0.000)	84.72 (0.000)
AIC	22915.99	22081.44	19261.65	15362.92
Two				
Coeff. (SD)	1.73 (2.63)	1.68 (2.71)	2.13 (3.39)	0.28 (5.13)
F (p-value)	0.03 (0.866)	0.01 (0.921)	0.37 (0.545)	0.18 (0.672)
AIC	22917.75	22083.22	19263.35	15364.95
Three				
Coeff. (SD)	2.69 (2.57)	2.95 (2.85)	2.87 (5.12)	-2.65 (12.57)
F (p-value)	0.19 (0.665)	0.19 (0.663)	0.00 (0.991)	0.75 (0.386)
AIC	22918.87	22084.31	19265.31	15366.69
Observations	2760	2532	2209	1760

Note: The cells contain the coefficient B_1 (robust standard errors) of equation 6 with controls, F test for the addition of polynomials of the running variable (p-value) -considering the first order polynomial the baseline-, and AIC. The bandwidth of 0.706 is the optimal bandwidth suggested by the bandwidth selector proposed by Calonico, Cattaneo and Titiunik (2014). Models include both an additional higher order polynomial and the interaction term between the instrumental variable and the running score at a higher order.
* p<0.1, **p<0.05, ***p<0.01

Table 24. Covariate balance above and below the ICE index threshold of the "autonomous" classification in 2012.

Outcomes	(1) Baseline	(2) With controls
Urban schools on t-1	0.00	0.02
	(0.04)	(0.04)
	2212	2212
Municipal school on t-1	-0.09	-0.05
	(0.07)	(0.06)
	2212	2212
Average enrollment on t-1	74.82	56.92
	(56.05)	(53.25)
	2212	2212
Enrollment 4th grade on t-1	3.83	3.11
	(4.62)	(4.39)
	2212	2212
Enrollment 8th grade on t-1	5.34	5.70
	(4.87)	(4.60)
	2212	2212
SES on t-1	-0.02	-0.11
	(0.10)	(0.09)
	2208	2208
Beneficiary students enrolled on t-1	27.07	15.77
	(17.39)	(11.32)
	2212	2212
Students with special needs enrolled on t-1	1.08	1.75
	(4.56)	(4.09)
	2212	2212

Note: Demographic variables listed in the first column are used as outcomes. The cells show resulting coefficient B_1 from equation 6 with and without covariates correspondingly (the outcome variable is removed from the list of covariates controlled for on column 2). The cells contain the main coefficient, the robust standard error on parenthesis, and the sample size. All outcomes are estimated using ordinary least squares. The schools included in the sample are those SEP schools that are classified using the classification formula, and that are within the 0.706 bandwidth.
* p<0.1, **p<0.05, ***p<0.01

Table 25. Impact of "autonomous" classification in 2012 on high-stakes outcomes.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	All schools		By pressure		By SES				By competition	
Outcomes	Baseline	With controls	Autonomous prior year	Not-autonomous prior year	Low SES	Mid Low SES	Mid SES	Mid- and High-SES	Low comp.	High comp.
4th grade Math on year t	2.82	0.81	3.47	1.29	1.08	1.35	2.28	-5.45	0.24	4.59
	(3.08)	(2.73)	(6.99)	(3.08)	(10.85)	(4.13)	(4.06)	(7.17)	(2.84)	(10.46)
	2209	2209	631	1578	230	1046	784	149	1973	236
4th grade Math on year t+1	4.95	2.72	1.47	4.03	-1.54	4.80	2.76	-4.53	2.27	6.55
	(3.10)	(2.80)	(6.70)	(3.16)	(9.29)	(4.41)	(4.28)	(6.19)	(2.94)	(9.09)
	2193	2193	629	1564	229	1036	779	149	1958	235
4th grade Lang. on year t	2.66	0.76	4.10	1.51	-9.96	5.04	-0.07	-7.88	0.18	13.41
	(2.74)	(2.36)	(5.65)	(2.71)	(9.53)	(3.61)	(3.60)	(5.16)	(2.48)	(8.11)
	2207	2207	631	1576	230	1045	783	149	1971	236
4th grade Lang. on year t+1	2.17	0.25	5.02	0.67	-12.80	1.87	1.78	-6.40	-0.27	8.29
	(2.73)	(2.36)	(5.58)	(2.69)	(9.59)	(3.67)	(3.51)	(4.51)	(2.48)	(7.18)
	2193	2193	629	1564	229	1036	779	149	1958	235
4th grade Science on year t	3.14	1.05	0.99	1.35	7.45	1.84	0.72	-9.11	0.80	1.79
	(2.77)	(2.22)	(5.95)	(2.46)	(8.67)	(3.49)	(3.13)	(5.93)	(2.30)	(8.30)
	2209	2209	631	1578	230	1046	784	149	1973	236
4th grade Science on year t+1	3.32	1.42	5.00	1.40	-7.24	3.02	2.38	-8.18	0.86	7.47
	(2.54)	(2.15)	(5.43)	(2.41)	(8.25)	(3.28)	(3.35)	(4.41)	(2.24)	(6.67)
	2186	2186	625	1561	226	1034	777	149	1953	233

Note: The table shows resulting coefficient B_1 from equation 6. Columns (2) to (10) include controls. The cells contain the main coefficient, the robust standard error on parenthesis, and the sample size. All outcomes are estimated using OLS. The schools included in the sample are those SEP schools that are classified using the classification formula, and that are within the 0.706 bandwidth. Columns (3) to (10) are estimations for subsamples according to level of pressure due to previous year classification, to SES level and to level of competition within the municipality of the school. A school is considered to be in a highly competitive municipality (High comp.) if the share of autonomous schools in a municipality is equal or greater than .2. Schools in low competitive municipalities (Low comp.) are all the other schools.

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

Table 26. Impact of "autonomous" classification in 2012 on low-stakes outcomes.

	(1)	(2)
Outcomes	Baseline	With controls
8th grade Math scores on year t+1	3.36	0.62
	(2.64)	(2.05)
	2154	2154
8th grade Language scores on year t+1	2.81	0.60
	(2.65)	(2.24)
	2150	2150
8th grade Science scores on year t+1	2.59	-0.26
	(2.52)	(1.89)
	2153	2153

Note: The table shows resulting coefficient $B_1$from equation 6 with and without controls correspondingly. The cells contain the main coefficient, the robust standard error on parenthesis, and the sample size. Models are estimated using OLS. The schools included in the sample are those SEP schools that are classified using the classification formula, and that are within the 0.706 bandwidth.
* p<0.1, **p<0.05, ***p<0.01

Table 27. Impact of "autonomous" classification in 2012 on school behavioral response.

	(1)	(2)	(3)	(4)
	All schools		School autonomous the prior year	Schools not-autonomous the prior year
Outcomes	Baseline	With controls	With controls	With controls
Tests taken on year t in 4th grade	19.42	1.64	1.21	1.46
	(12.53)	(1.42)	(3.34)	(1.62)
	2156	2156	614	1542
Tests taken on year t+1 in 4th grade	18.65	0.84	-11.64	3.77
	(12.30)	(4.05)	(9.16)	(4.60)
	2150	2150	615	1535
Students with special needs on year t	14.57	12,13	-6.52	23.16
	(14.08)	(13.86)	(44.11)	(12.98)
	2212	2212	632	1580
Students with special needs on year t+1	-0.32	-0.20	-10.33	3.91
	(5.92)	(5.21)	(13.05)	(5.70)
	2198	2198	630	1568
Students retained in 3th grade on year t	-0.04	-0.17	-0.16	-0.26
	(0.32)	(0.28)	(0.67)	(0.31)
	2212	2212	632	1580
Students retained in 3th grade on year t+1	0.09	-0.06	-0.38	-0.02
	(0.32)	(0.29)	(0.53)	(0.35)
	2198	2198	630	1568
Students approved in 3th grade on year t	5.47	-0.01	-3.47	1.27
	(4.14)	(1.33)	(3.03)	(1.54)
	2212	2212	632	1580
Students approved in 3th grade on year t+1	7.52	2.03	2.50	1.75
	(4.16)	(1.43)	(3.67)	(1.58)
	2198	2198	630	1568
Students switching schools in 3th grade on year t	-0.41	-0.51	-0.29	-0.68
	(0.45)	(0.39)	(0.90)	(0.46)
	2212	2212	632	1580
Students switching schools in 3th grade on year t+1	0.42	0.23	0.64	0.21
	(0.45)	(0.39)	(0.93)	(0.44)
	2198	2198	630	1568
Proportion of students scoring above 250 on year t	0.02	0.00	0.01	0.01
	(0.02)	(0.02)	(0.05)	(0.02)
	2148	2148	612	1536
Proportion of students scoring above 250 on year t+1	0.03	0.01	0.03	0.02
	(0.02)	(0.02)	(0.05)	(0.02)
	2156	2156	619	1537

Table 27, Continued.

	(1)	(2)	(3)	(4)
	All schools		School autonomous the prior year	Schools not-autonomous the prior year
Outcomes	Baseline	With controls	With controls	With controls
Proportion of students scoring above 300 on year t	0.03	0.02	0.00	0.02
	(0.01)	(0.01)	(0.03)	(0.01)
	2148	2148	612	1536
Proportion of students scoring above 300 on year t+1	0.02	0.00	0.01	0.00
	(0.01)	(0.01)	(0.03)	(0.01)
	2156	2156	619	1537
ICE index on year t	0.09	0.00	0.09	-0.02
	(0.07)	(0.04)	(0.09)	(0.04)
	2184	2184	624	1560
ICE index on year t+1	0.07	0.00	0.08	-0.00
	(0.06)	(0.04)	(0.11)	(0.05)
	2157	2157	615	1542
SNED initiative on t	-2.16	-1.89	10.16	-4.34
	(3.24)	(3.14)	(7.03)	(3.44)
	2176	2176	622	1554
SNED initiative on t+1	-1.65	-1.36	10.95	-4.03
	(3.25)	(3.15)	(7.04)	(3.46)
	2150	2150	613	1537
SNED improvement on t	0.93	0.87	-0.32	0.02
	(1.63)	(1.63)	(3.07)	(2.06)
	2176	2176	622	1554
SNED improvement on t+1	0.86	0.82	-0.22	-0.01
	(1.64)	(1.64)	(3.09)	(2.07)
	2150	2150	613	1537
SNED integration on t	-2.53	-2.62	8.09	-4.63
	(3.23)	(3.20)	(7.16)	(3.55)
	2176	2176	622	1554
SNED integration on t+1	-2.19	-2.36	8.81	-4.55
	(3.24)	(3.21)	(7.18)	(3.57)
	2150	2150	613	1537

Table 27, Continued.

	(1)	(2)	(3)	(4)
	All schools		School autonomous the prior year	Schools not-autonomous the prior year
Outcomes	Baseline	With controls	With controls	With controls
Approval rates on t[#]	-0.00	-0.00	0.02	-0.00
	(0.01)	(0.01)	(0.01)	(0.01)
	2184	2184	624	1560
Approval rates on t+1[#]	-0.01	-0.01	0.01	-0.01
	(0.01)	(0.01)	(0.01)	(0.01)
	2158	2158	615	1543
Retention rates on t[#]	-0.01	-0.01*	0.00	-0.01*
	(0.00)	(0.00)	(0.01)	(0.00)
	2184	2184	624	1560
Retention rates on t+1[#]	0.00	-0.01*	-0.00	-0.01
	(0.00)	(0.00)	(0.01)	(0.00)
	2158	2158	615	1543
Teacher evaluation t [@]	0.01	0.01	0.00	0.01
	(0.03)	(0.03)	(0.07)	(0.04)
	1330	1330	403	927
Teacher evaluation t+1 [@]	-0.00	-0.00	0.07	-0.02
	(0.03)	(0.03)	(0.07)	(0.04)
	1321	1321	402	919

Note: The table shows resulting coefficient B_1 from equation 6 with and without covariates correspondingly. The cells contain the main coefficient, the robust standard error on parenthesis, and the sample size. All outcomes are estimated using OLS. The schools included in the sample are those SEP schools that are classified using the classification formula, and that are within the 0.706 bandwidth. The two columns at the right divide the sample on whether the school was classified as autonomous the previous year or not. [#]These models do not include enrollment rates as a control variable because it is also a denominator of the outcome. [@] Sample restricted to only public schools, because private voucher schools do not have teacher evaluations. * p<0.1, **p<0.05, ***p<0.01

Table 28. Impact of "autonomous" classification in 2012 on student body composition.

	(1)	(2)	(3)	(4)
	All schools		School autonomous the prior year	Schools not-autonomous the prior year
Outcomes	Baseline	With controls	With controls	With controls
Total school enrollment on year t	74.09	3.92	-87.57	17.81
	(52.76)	(24.54)	(65.54)	(25.24)
	2212	2212	632	1580
Total school enrollment on year t+1	77.29	4.61	-63.52	12.50
	(52.55)	(24.59)	(63.22)	(26.02)
	2198	2198	630	1568
School enrollment in 1st grade on year t	8.63*	3.06	-2.94	3.26
	(4.37)	(1.65)	(4.12)	(1.87)
	2212	2212	632	1580
School enrollment in 1st grade on year t+1	7.52	1.48	-1.30	2.05
	(4.58)	(1.89)	(3.78)	(2.24)
	2198	2198	630	1568
School enrollment in 7th grade on year t	6.03	1.63	-1.97	1.60
	(4.39)	(2.24)	(4.77)	(2.60)
	2212	2212	632	1580
School enrollment in 7th grade on year t+1	6.75	1.92	0.38	1.30
	(4.53)	(2.23)	(4.72)	(2.57)
	2198	2198	630	1568
Enrollment fees on year t+1 @	0.08	0.06	-0.04	0.06
	(0.12)	(0.11)	(0.29)	(0.13)
	866	866	226	640
Monthly fees on year t+1 @	0.16	0.09	-0.13	0.07
	(0.23)	(0.19)	(0.49)	(0.22)
	863	863	225	638
Changes in SES level of 4th grade from year t-1 to t#	-0.11	-0.10	-0.09	-0.11
	(0.06)	(0.06)	(0.14)	(0.07)
	2160	2160	620	1540
Changes in SES level of 4th grade from year t-1 to t+1#	0.09	0.10	0.00	0.10
	(0.07)	(0.07)	(0.15)	(0.08)
	2170	2170	627	1543
Beneficiary students enrolled on year t	25.52	12.48	-0.95	14.80
	(19.16)	(9.31)	(23.52)	(9.95)
	2212	2212	620	1580
Beneficiary students enrolled on year t+1	37.79	18.09	17.54	19.43
	(23.81)	(11.87)	(29.56)	(12.72)
	2198	2198	630	1568

Note: The table shows resulting coefficient B_1 from equation 6 with and without covariates correspondingly. Cells contain the main coefficient (robust standard error) and the sample size. All outcomes are estimated using OLS. Schools included in the sample are those SEP schools that are classified using the classification formula, and that are within the 0.706 bandwidth. Column (3) and (4) divide the sample on whether the school was classified as autonomous the previous year or not. @ Sample restricted to only private voucher schools, because primary public schools are not allowed to charge fees on top of the voucher. # These models do not control for the SES level at the baseline to avoid a spurious correlation. * p<0.1, **p<0.05, ***p<0.01

Table 29. Impact of "autonomous" classification in 2012 on teacher body composition.

	(1)	(2)	(3)	(4)
	All schools		School autonomous the prior year	Schools not-autonomous the prior year
Outcomes	Baseline	With controls	With controls	With controls
Teachers teaching on year t	2.37	0.33	-2.11	1.10
	(1.90)	(1.24)	(3.38)	(1.33)
	2212	2212	632	1580
Teachers teaching on year t+1	2.57	0.39	-0.33	0.96
	(1.93)	(1.26)	(3.40)	(1.35)
	2198	2198	630	1568
Teaching hours on year t	82.70	9.99	-96.42	40.93
	(68.43)	(43.10)	(118.70)	(45.03)
	2212	2212	632	1580
Teaching hours on year t+1	81.39	2.45	-38.76	29.01
	(70.52)	(45.26)	(121.89)	(47.33)
	2198	2198	630	1568

Note: The table shows resulting coefficient B_1 from equation 6 with and without covariates correspondingly. The cells contain the main coefficient, the robust standard error on parenthesis, and the sample size. All outcomes are estimated using OLS. The schools included in the sample are those SEP schools that are classified using the classification formula, and that are within the 0.706 bandwidth. The two columns at the right divide the sample on whether the school was classified as autonomous the previous year or not.
* p<0.1, **p<0.05, ***p<0.01

Table 30. First stage models for schools being classified or not "in recovery" classification in 2012 & 2013 rounds pooled.

	(1) Baseline	(2) With controls
Below ICE $_{t-1}$	0.90***	0.90***
	(0.03)	(0.024)
ICE $_{t-1}$	-0.237***	-0.242***
	(0.00)	(0.66)
Below ICE $_{t-1}$ * ICE $_{t-1}$	0.197***	0.196***
	(0.07)	(0.073)
Observations	1054	1054
R^2	0.903	0.905
F	1121.89	1075.12

Notes: The table shows resulting coefficients from equation 7 estimated with a linear probability model. The cells contain the main coefficients and the robust standard errors in parenthesis. The schools included are those SEP schools that are classified according to the classification formula, and that are within the 0.467 bandwidth. Covariates control for whether the school is in an urban area, the type of school, schools SES level, and the enrollment rates in 4th grade. The models also include fixed effects per year.
* p<0.1, **p<0.05, ***p<0.01

Table 31. Specification check of RD estimate of receiving a school classification of "in recovery" on mathematics outcomes in 4th grade on year t. Rounds 2012 & 2013 pooled.

	(1)	(2)	(3)	(4)
Bandwidth:	[-1 ; 1]	[-.7 ; .7]	[-.467 ; .467]	[-.3 ; .3]
Polynomial of order:				
One				
Coeff. (SD)	4.65* (1.99)	4.36 (2.25)	4.75 (2.80)	5.23 (3.69)
F (p-value)	50.54 (0.000)	21.10 (0.000)	8.71 (0.000)	2.76 (0.005)
AIC	23582.53	15254.86	9239.63	5402.06
Two				
Coeff. (SD)	3.85 (2.34)	3.75 (2.51)	4.41 (2.93)	4.67 (3.76)
F (p-value)	0.03 (0.853)	0.12 (0.734)	0.01 (0.907)	0.27 (0.601)
AIC	23583.96	15256.70	9241.59	5404.20
Three				
Coeff. (SD)	4.28 (2.64)	4.63 (3.05)	4.88 (3.86)	8.32 (5.37)
F (p-value)	0.04 (0.846)	0.27 (0.603)	1.07 (0.301)	0.06 (0.809)
AIC	23585.68	15258.32	9243.79	5405.10
Observations	2693	1741	1054	613

Note: The cells contain the coefficient ϑ_1 (robust standard errors) of equation 8 with controls and year fixed effects, F test for the addition of polynomials of the running variable (p-value) -considering the first order polynomial the baseline-, and AIC. The bandwidth of 0.467 is the optimal bandwidth suggested by the bandwidth selector proposed by Calonico, Cattaneo and Titiunik (2014). Models include both an additional higher order polynomial and the interaction term between the instrumental variable and the running score at a higher order.
* p<0.1, **p<0.05, ***p<0.01

Table 32. Covariate balance above and below the ICE index threshold of the "in recovery" classification. Rounds 2012 & 2013 pooled.

Outcomes	(1) Baseline	(2) With controls
Urban schools on t-1	-0.01	-0.01
	(0.03)	(0.03)
	1063	1063
Municipal school on t-1	0.03	0.05
	(0.07)	(0.06)
	1063	1063
Average enrollment on t-1	-7.95	-8.32
	(33.95)	(33.32)
	1063	1063
Enrollment 4th grade on t-1	-1.28	-1.29
	(3.23)	(3.22)
	1063	1063
Enrollment 8th grade on t-1	-4.76	-5.07
	(3.58)	(3.46)
	1063	1063
SES on t-1	-0.07	-0.05
	(0.08)	(0.08)
	1058	1058
Beneficiary students enrolled on t-1	-14.18	0.72
	(16.68)	(7.57)
	1063	1063
Students with special needs enrolled on t-1	-0.74	0.80
	(7.64)	(7.58)
	1063	1063

Note: Demographic variables listed in the first column are used as outcomes. The cells show resulting coefficient ϑ_1 from equation 8 with and without covariates correspondingly (the outcome variable is removed from the list of covariates controlled for on column 2). The cells contain the main coefficient, the robust standard error on parenthesis, and the sample size. All outcomes are estimated using ordinary least squares. The schools included in the sample are those SEP schools that are classified using the classification formula, and that are within the 0.467 bandwidth.

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

Table 33. Impact of "in recovery" classification on high-stakes outcomes. Rounds 2012 & 2013 pooled.

Outcomes	(1) Baseline	(2) With controls
4th grade Math on year t	4.65	4.75
	(2.82)	(2.80)
	1054	1054
4th grade Math on year t+1#	6.75	6.80
	(4.35)	(4.30)
	524	524
4th grade Lang. on year t	1.55	1.52
	(2.55)	(2.53)
	1053	1053
4th grade Lang. on year t+1#	7.02	7.08
	(3.84)	(3.73)
	524	524
4th grade Science on year t	1.47	1.53
	(2.31)	(2.28)
	1050	1050
4th grade Science on year t+1#	4.32	3.75
	(3.14)	(3.03)
	520	520

Note: The table shows resulting coefficient ϑ_1 from equation 8. Column (2) includes controls. The cells contain the main coefficient, the robust standard error on parenthesis, and the sample size. All outcomes are estimated using OLS. The schools included in the sample are those SEP schools that are classified using the classification formula, and that are within the 0.467 bandwidth. #Only considers outcomes for the classification of year 2012, those are of the test in 2013. That is why the sample size is half the size.
* p<0.1, **p<0.05, ***p<0.01

Table 34. Impact of "in recovery" classification on low-stakes outcomes. Rounds 2012 & 2013 separate.

	(1)	(2)	(3)	(4)
	2012 classification		2013 classification	
Outcomes	Baseline	With controls	Baseline	With controls
8th grade Math scores on year t	-	-	-3.99	-3.53
			(2.48)	(2.46)
			510	510
8th grade Language scores on year t	-	-	-0.61	-1.00
			(3.10)	(2.94)
			509	509
8th grade Science scores on year t	-	-	-2.58	-1.87
			(2.46)	(2.20)
			510	510
8th grade Math scores on year t+1	7.52**	7.04*	-	-
	(2.90)	(2.69)		
	521	521		
8th grade Language scores on year t+1	5.09	5.52	-	-
	(3.83)	(3.65)		
	518	518		
8th grade Science scores on year t+1	4.18	3.82	-	-
	(2.73)	(2.51)		
	520	520		

Note: The table shows resulting coefficient ϑ_1 from equation 8 with and without controls correspondingly. The cells contain the main coefficient, the robust standard error on parenthesis, and the sample size. Models are estimated using OLS. The schools included in the sample are those SEP schools that are classified using the classification formula, and that are within the 0.467 bandwidth. Given the peculiarities of the evaluation calendar, there are no test scores for 8th grade on year 2012. For this reason, the analysis could not be performed for both rounds pooled.
* p<0.1, **p<0.05, ***p<0.01

Table 35. Impact of "in recovery" classification on school behavioral response. Rounds 2012 & 2013 pooled.

Outcomes	(1) Baseline	(2) With controls
Tests taken on year t in 4th grade	-8.29	-0.28
	(7.81)	(1.64)
	1029	1029
Students with special needs on year t	5.50	6.83
	(8.26)	(8.19)
	1063	1063
Students retained in 3th grade on year t	0.15	0.29
	(0.34)	(0.31)
	1063	1063
Students approved in 3th grade on year t	-2.83	-0.23
	(2.69)	(1.29)
	1063	1063
Students switching schools in 3th grade on year t	-0.26	0.03
	(0.62)	(0.51)
	1063	1063
Proportion of students scoring above 250 on year t	0.02	0.02
	(0.02)	(0.02)
	1030	1030
Proportion of students scoring above 300 on year t	0.01	0.01
	(0.01)	(0.01)
	1030	1030
ICE index on year t	0.03	0.03
	(0.04)	(0.04)
	1050	1050
SNED initiative on t	5.61	5.65
	(5.13)	(4.85)
	1044	1044
SNED improvement on t	0.70	0.79
	(1.58)	(1.57)
	1044	1044
SNED integration on t	3.97	3.86
	(4.55)	(4.37)
	1044	1044
Approval rates on t#	-0.00	-0.00
	(0.01)	(0.01)
	1050	1050

Table 35, Continued.

Outcomes	(1) Baseline	(2) With controls
Retention rates on t[#]	-0.01**	-0.01**
	(0.00)	(0.00)
	1050	1050
Teacher evaluation t [@]	0.04	0.04
	(0.03)	(0.03)
	731	731

Note: The table shows resulting coefficient ϑ_1 from equation 8 with and without covariates correspondingly. The cells contain the main coefficient, the robust standard error on parenthesis, and the sample size. All outcomes are estimated using OLS. The schools included in the sample are those SEP schools that are classified using the classification formula, and that are within the 0.467 bandwidth. [#]These models do not include enrollment rates as a control variable because it is also a denominator of the outcome. [@] Sample restricted to only public schools, because private voucher schools do not have teacher evaluations.
* p<0.1, **p<0.05, ***p<0.01

Table 36. Impact of "in recovery" classification on student body composition. Rounds 2012 & 2013 pooled.

Outcomes	(1) Baseline	(2) With controls
Total school enrollment on year t	-22.20	2.40
	(31.29)	(21.00)
	1063	1063
School enrollment in 1st grade on year t	-3.49	-0.64
	(3.03)	(1.52)
	1063	1063
School enrollment in 7th grade on year t	-5.06	-2.54
	(3.26)	(1.75)
	1063	1063
Changes in SES level of 4th grade from year t-1 to t#	-0.10	-0.03
	(0.11)	(0.10)
	520	520
Beneficiary students enrolled on year t	-16.83	0.10
	(17.89)	(7.65)
	1063	1063

Note: The table shows resulting coefficient ϑ_1 from equation 8 with and without covariates correspondingly. The cells contain the main coefficient, the robust standard error on parenthesis, and the sample size. All outcomes are estimated using OLS. The schools included in the sample are those SEP schools that are classified using the classification formula, and that are within the 0.467 bandwidth. # This model does not control for the SES level at the baseline to avoid a spurious correlation.
* p<0.1, **p<0.05, ***p<0.01

Table 37. Impact of "in recovery" classification on teacher body composition. Rounds 2012 & 2013 pooled.

Outcomes	(1) Baseline	(2) With controls
Teachers teaching on year t	0.12	0.67
	(1.38)	(1.16)
	1062	1062
Teaching hours on year t	-4.82	19.62
	(49.98)	(39.06)
	1062	1062

Note: The table shows resulting coefficient ϑ_1 from equation 8 with and without covariates correspondingly. The cells contain the main coefficient, the robust standard error on parenthesis, and the sample size. All outcomes are estimated using OLS. The schools included in the sample are those SEP schools that are classified using the classification formula, and that are within the 0.467 bandwidth.
* p<0.1, **p<0.05, ***p<0.01

Figures

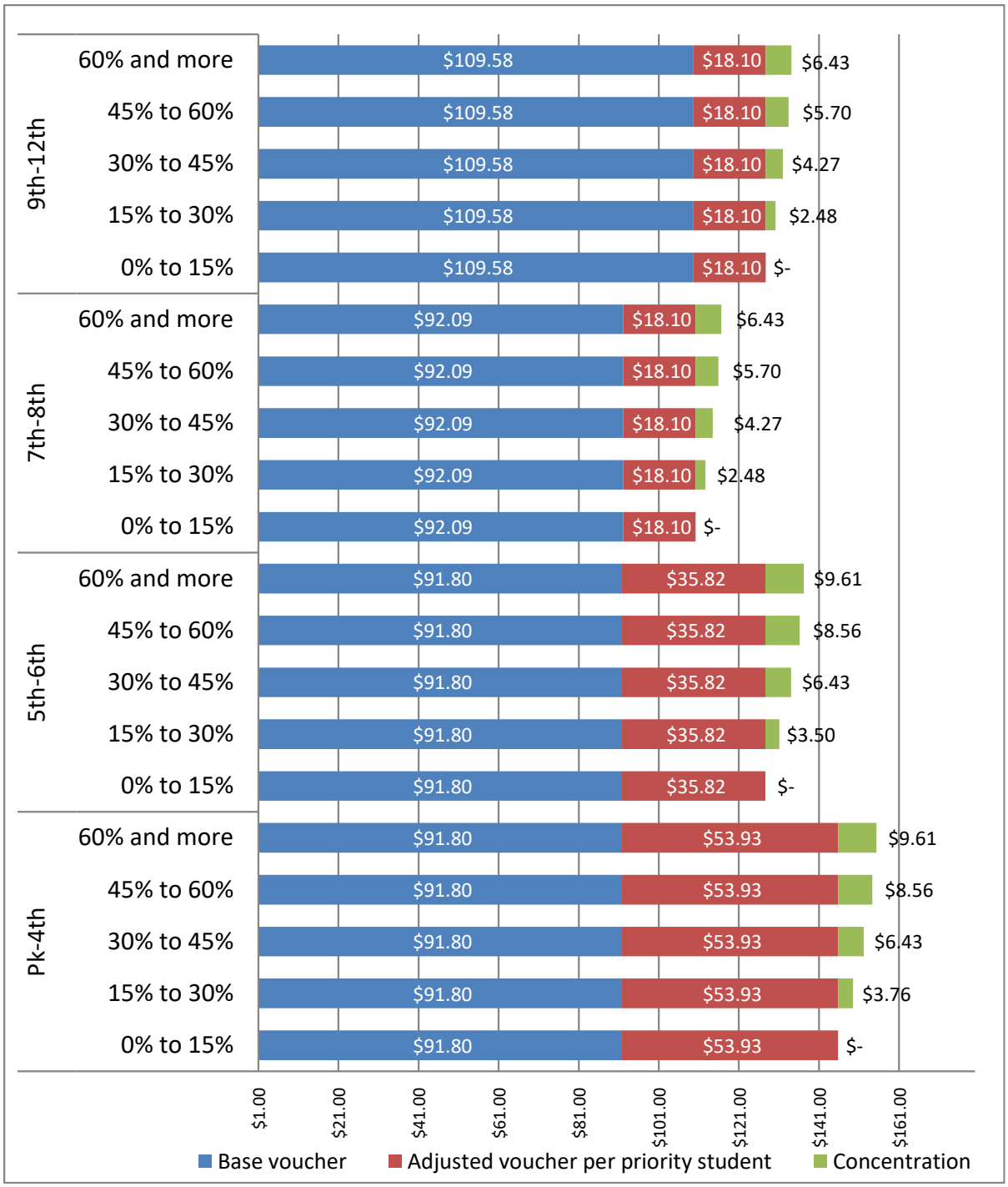


Figure 1. Monthly voucher values per grade and concentration levels.

Note: Data from the Law 20550 (from 2011), Law 20501 (from 2011), and value of the vouchers starting December 2014. Values in 2015 USD.

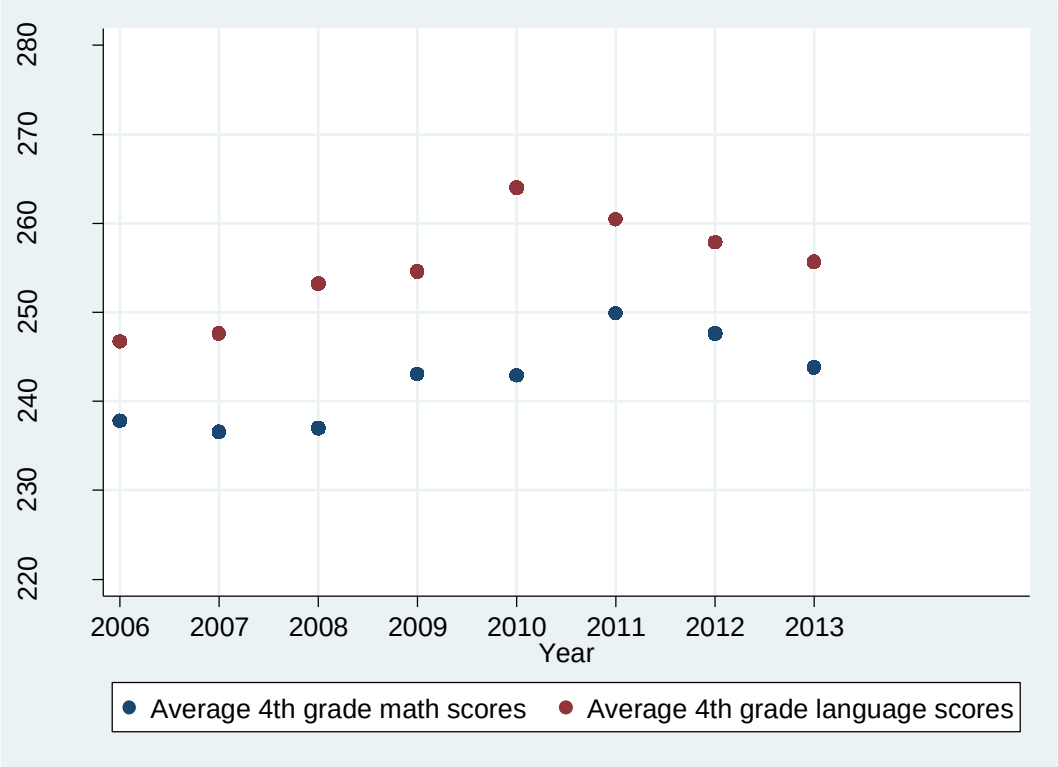


Figure 2. 4th grade national math and language test scores from 2006 to 2013.

Note: Data from *Agencia de Calidad de la Educación*. Schools included are those that have 4th grade with 20 or more students taking the tests.

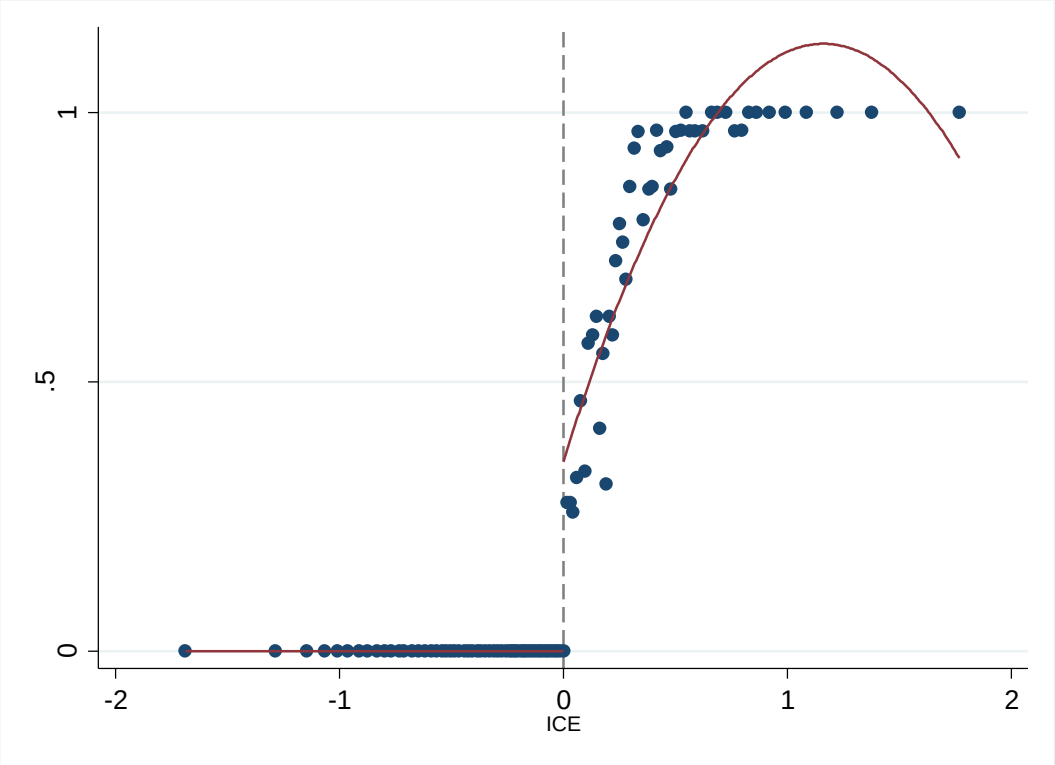
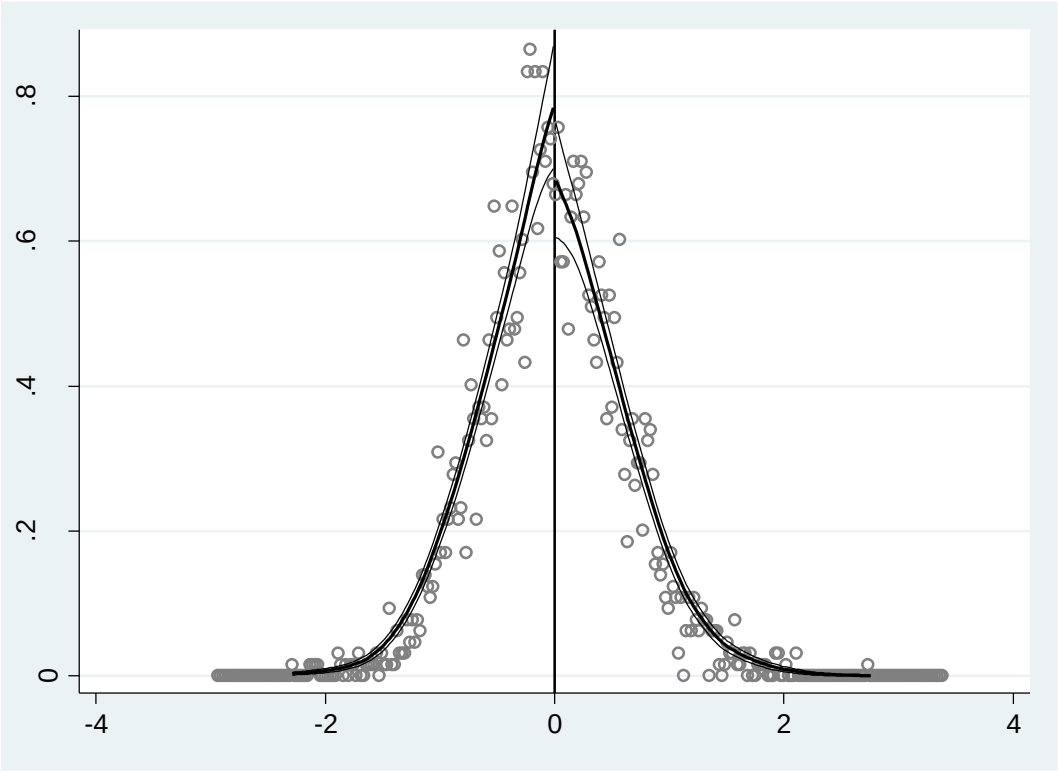


Figure 3. Proportion of schools classified as "autonomous" on 2012 by ICE index relative to median of the SES group.

Note: Each circle represents the proportion of schools classified as "autonomous" within equal-sized ICE-index bins relative to the median of the SES group (the dotted line). The fitted line results from the underlying data. The plot is based on 2,901 school observations.

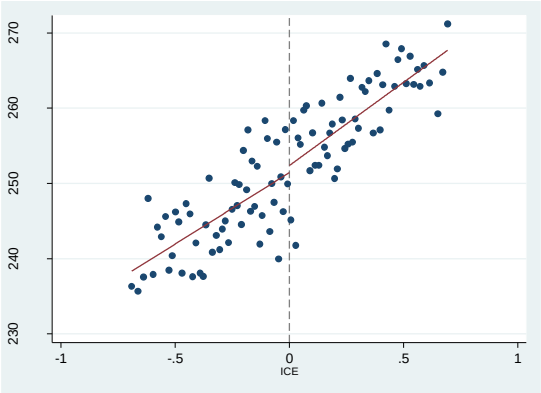


Log difference in height=-0.14; *p*=.08; n=2901

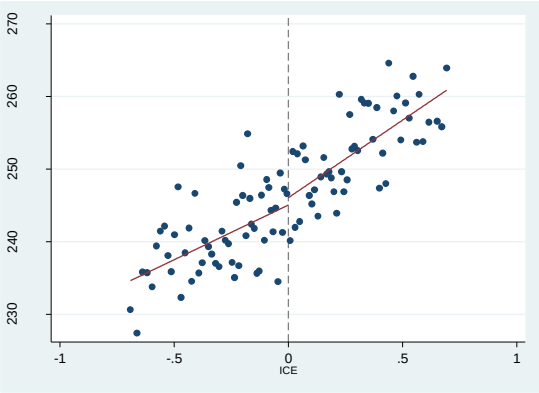
Figure 4. McCrary density test. Density of ICE index for schools in 2012 at the "autonomous" threshold centered at the median of the SES groups.

Note: Schools included are all those elementary schools that have adopted the SEP law, that are classified using the classification formulae. Data from 2012.

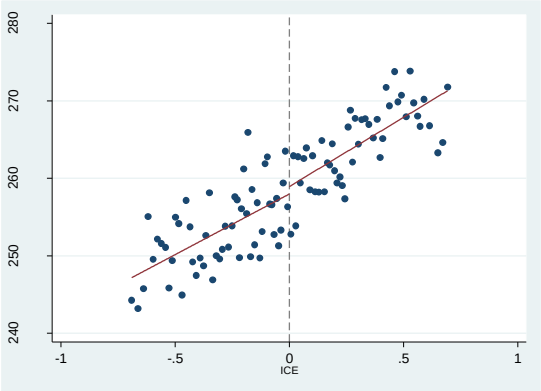
Panel A. Math on year t



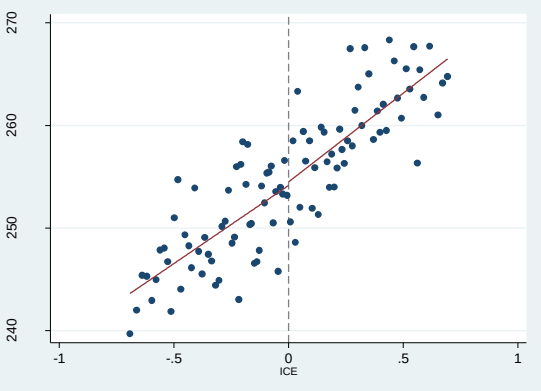
Panel D. Math on year t+1



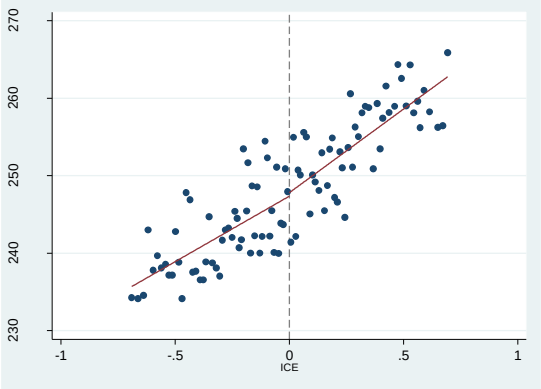
Panel B. Language on year t



Panel E. Language on year t+1



Panel C. Science on year t



Panel F. Science on year t+1

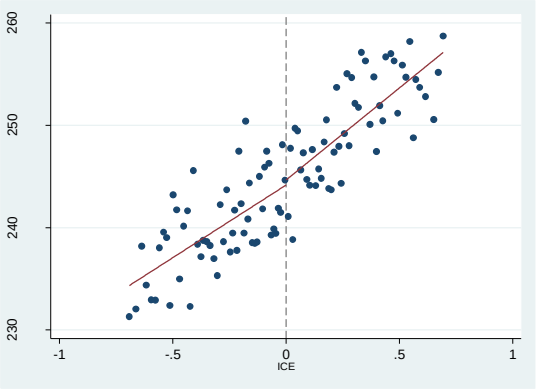


Figure 5. Selected outcomes by ICE index at the "autonomous" cutoff.

Note: Each circle represents the SIMCE scores of the schools by the ICE index relative to the median of the SES group. Schools are binned into 100 equal-sized bins. Fitted lines do not control for covariates.

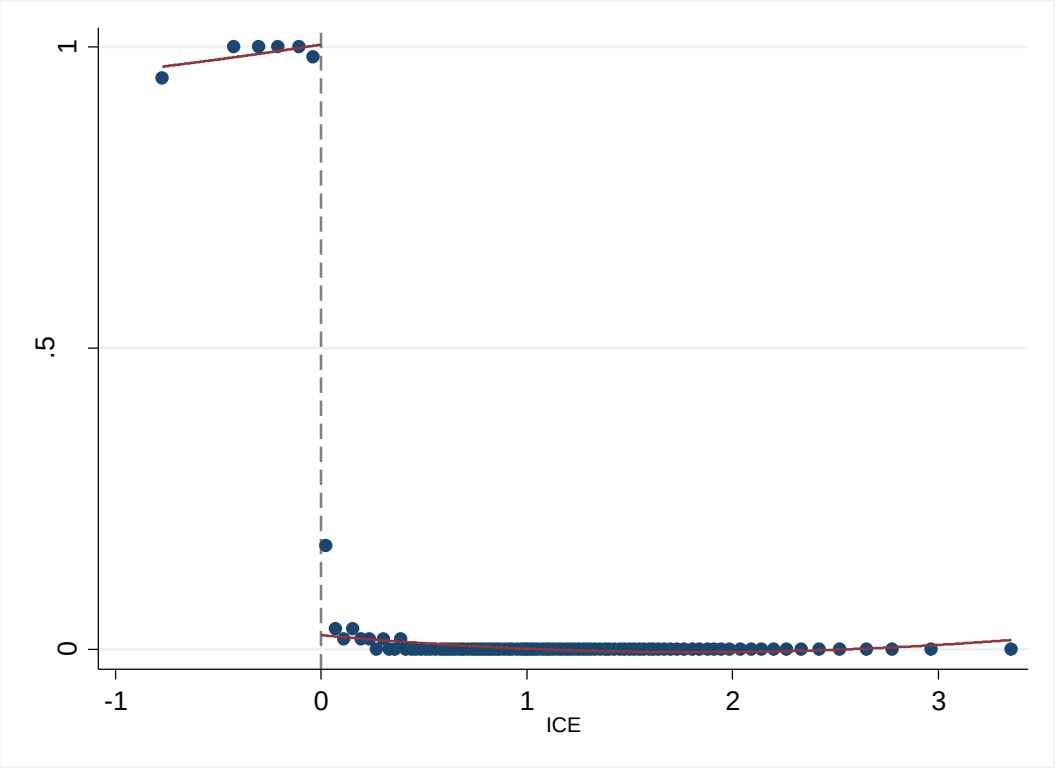
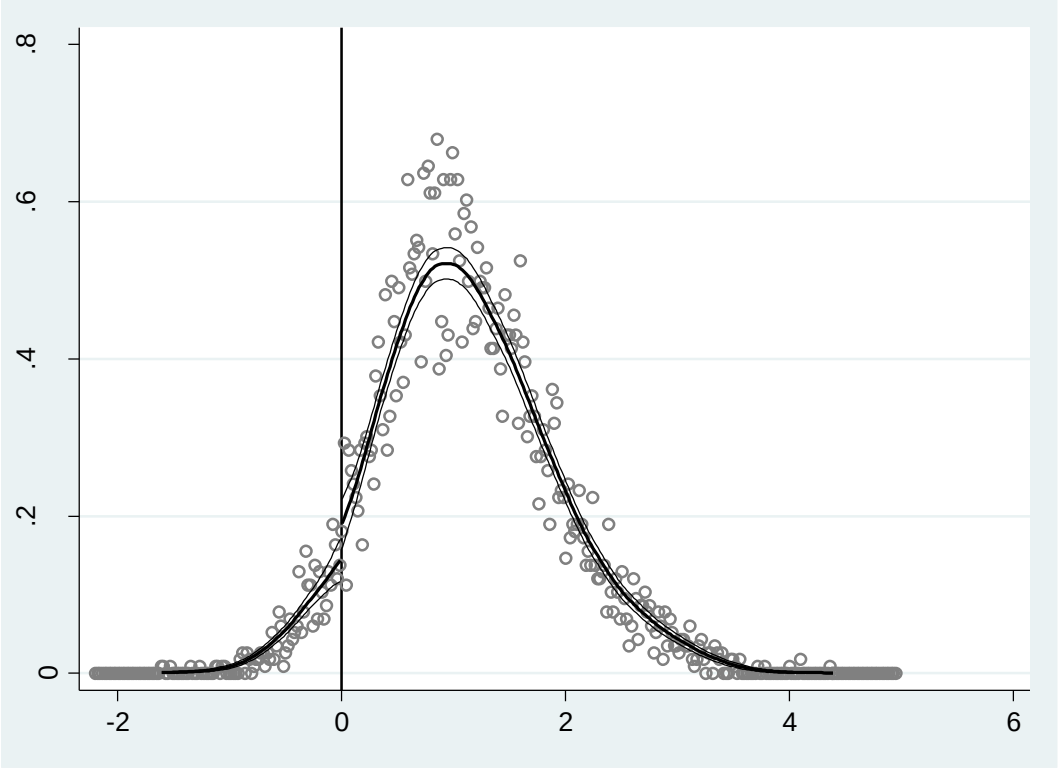


Figure 6. Proportion of schools classified as "in recovery" on 2012 and 2013 by ICE index.

Note: Each circle represents the proportion of schools classified as "in recovery" within equal-sized ICE-index bins relative to the median of the SES group (the dotted line). The fitted line results from the underlying data. The plot is based on 5,763 school observations.

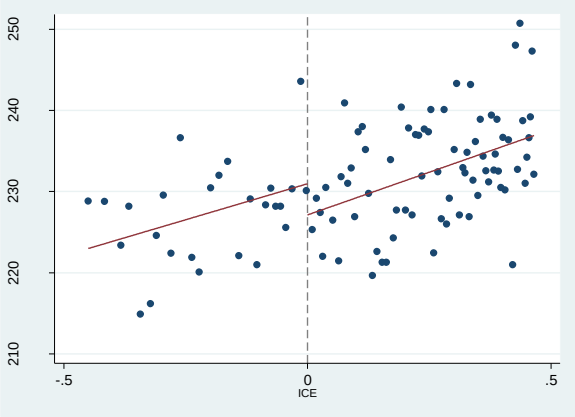


Log difference in height=0.247; *p*=.13; n=5763

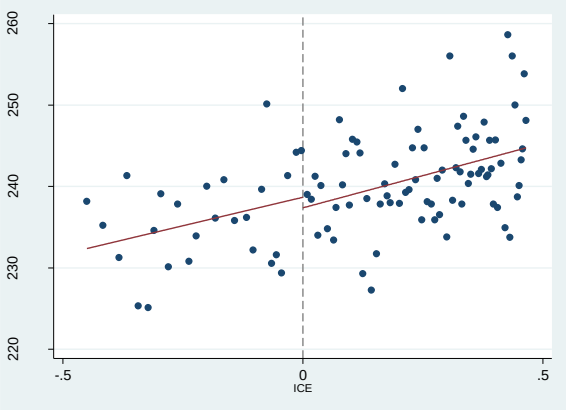
Figure 7. McCrary density test. Density of ICE index for schools in 2012 and 2013 at the "in recovery" threshold.

Note: Schools included are all those elementary schools that have adopted the SEP law, that are classified using the classification formulae. Data from 2012 & 2013.

Panel A. Math on year t



Panel B. Language on year t



Panel C. Science on year t

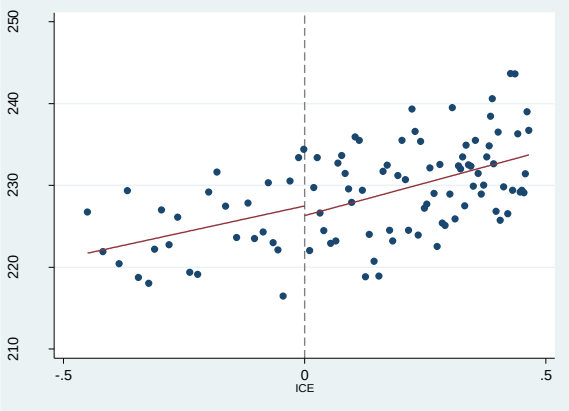


Figure 8. Selected outcomes by ICE index at the "in recovery" cutoff. Rounds 2012 & 2013 pooled.

Note: Each circle represents the SIMCE scores of the schools by the ICE index relative to the cutoff. Schools are binned into 100 equal-sized bins. Fitted lines do not control for covariates. Data from years 2012 and 2013.

References

Allen, R., & Burgess, S. (2012). How should we treat under-performing schools? A regression discontinuity analysis of school inspections in England. Working paper 12/287. *The Centre for Market and Public Organization (CMPO).*

Allende, C. (2012). The impact of information on academic achievement and school choice: Evidence from Chilean "traffic lights". Master thesis. Instituto de Economia. Pontificia Universidad Catolica de Chile.

Anand, P., Mizala, A., & Repetto, A. (2009). Using school scholarships to estimate the effect of private education on the academic achievement of low-income students in Chile. *Economics of Education Review, 28*, 370-381.

Andrabi, T., Das, J., & Khawaja, A. (2014). Report cards: The impact of providing school and child test scores on educational markets. *HKS Working Paper* No. RWP14-052.

Angrist, J., & Pischke, J. (2009). *Mostly harmless econometrics: An empiricist's companion.* Princeton, NJ: Princeton University Press.

Baker, G. (2002). Distortion and risk in optimal incentive contracts. *The Journal of Human Resources, 37* (4). 728-751.

Baker, E. L., & Linn, R. L. (2004). Validity issues for accountability systems. In S. Fuhrman and R. Elmore (Eds.), *Redesigning accountability systems for education*. New York: Teachers College Press.

Barra, D. (2013). Evaluación del proceso de implementación de la subvención escolar preferencial. Tesis para optar al grado de Magister en Gestión y Políticas Públicas. Universidad de Chile: Facultad de Ciencias Físicas y Matemáticas.

Bases de Datos de la Agencia de Calidad de la Educación [2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015]. Santiago, Chile.

Baumeister, R., Bratslavsky, E., Finkenauer, C., & Vohs, K. (2001). Bad is stronger than good. *Review of General Psychology, 5*(4), 323-370.

Bush, J., Hough, H., & Kirst, M. (2017). How should states design their accountability systems? *Education Next, 17*(1). URL: educationnext.org

Calonico, S., Cattaneo, M. D., & Titiunik, R. (2014). Robust data-driven inference in the regression-discontinuity design. *Stata Journal, 14*(4), 909-946.

Carnoy, M., & Loeb, S. (2002). Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis, 24* (4), 305-331.

Carpenter-Huffman, P., Hall, G., & Sumner, G. (1975). *Change in education: Insights from performance contracting.* Massachusetts: Ballinger Publishing Company.

Chay, K. Y., McEwan, P. J., & Urquiola, M. (2005). The central role of noise in evaluating interventions that use test scores to rank schools. *The American Economic Review, 95*(4), 1237-1258.

Chiang, H. (2009). How accountability pressure on failing schools affects student achievement. *Journal of Public Economics, 93*, 1045-1057.

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

Clotfelter, C., Ladd, H., Vigdor, J., & Wheeler, J. (2006). High-poverty schools and the distribution of teachers and principals. *North Carolina Law Review, 85*, 1345-1380.

Cohodes, S. R., & Goodman, J. S. (2014). Merit aid, college quality, and college completion: Massachusetts' Adams scholarship as an in-kind subsidy. *American Economic Journal: Applied Economics, 6(*4), 251-285.

Contraloria General de la Republica (2014). Informe final 87-14 Subsecretaria de educación. Transferencias efectuadas con cargo a la subvención escolar preferencial. Noviembre 2014. URL: https://www.contraloria.cl/SicaProd/SICAv3-BIFAPortalCGR/faces/detalleInforme?docIdcm=230d6cac753ca1d9303a2c111fc71424&_adf.ctrl-state=mprkpvjfo_7

Correa, J., Inostroza, D., Parro, F., Reyes, L., & Ugarte, G. (2013). The effects of vouchers on academic achievement: Evidence from Chile's conditional voucher program. Gobierno de Chile, Ministerio de Hacienda.

Craig, S., Imberman, S., & Purdue, A. (2013). Does it pay to get an A? School resource reallocations in response to accountability ratings. *Journal of Urban Economics, 73*, 30-42.

Cullen J. B., & Reback, R. (2006). Tinkering toward accolades: School gaming under performance accountability system. In T. Gronberg and D. Jansen (2006), *Advances in Applied Microeconomics: Improving school accountability: check-ups or choice.* The Netherlands: Elsevier.

Deming, D., Cohodes, S., Jennings, J., & Jencks, C. (2013). School accountability, postsecondary attainment and earnings. Working paper 19444. .*National Bureau of Economic Research.*

Deming, D. J., & Figlio, D. (2016). Accountability in US Education: Applying Lessons from K–12 Experience to Higher Education. *The Journal of Economic Perspectives, 30*(3), 33-55.

Dixit, A. (2002). Incentives and organizations in the public sector: An interpretative review. *The Journal of Human Resources, 37* (4), 696-727.

Driver, C. (2003). Towards an economic model of school principal accountability. Dissertation submitted in partial fulfilment of the requirements of the degree of doctor of philosophy to the School of Education and the Committee on Graduate Studies of Stanford University.

Elacqua, G., Martínez, M., Santos, H., & Urbina, D. (2015). Short-run effects of accountability pressures on teacher policies and practices in the voucher system in Santiago, Chile. *School Effectiveness and School Improvement* (Online, September 2015).

Elacqua, G., Mosqueira, Ú., & Santos, H. (2009). La toma de decisiones de un sostenedor; Análisis a partir de la ley SEP. *En Foco Educación*, (1). URL: http://www.facso.uchile.cl/psicologia/epe/_documentos/GT_cultura_escolar_politica_educativa/recursos%20bibliograficos/articulos%20sep/elacquaetal(2009)decisionessep.pdf

Feng, L., Figlio, D., & Sass, T. (2010). School accountability and teacher mobility. NBER Working paper No. 16070.

Fernandez, S. (2009). Understanding contracting performance: An empirical analysis. *Administration and Society, 41* (1), 67-100.

Ferris, J. (1992). School-based decision making: A principal-agent perspective. *Educational Evaluation and Policy Analysis, 14*(4), 333-346.

Figlio, D., & Ladd, H. (2008). School accountability and student achievement. In H. Ladd and E. Fiske (Eds.), *Handbook of Research in Education Finance and Policy*. New York: Routledge.

Figlio, D., & Ladd, H. (2015). School accountability and student achievement. In H. Ladd and M. Goertz (Eds.), Handbook of Research in Education and Policy. New York: Routledge.

Figlio, D. & Loeb, S. (2011). School accountability. In E. Hanushek, S. Machin and L. Woessmann (Eds.), *Handbooks in Economics, Vol3*, The Netherlands: Elsevier.

Figlio, D., & Rouse, C. (2006). Do accountability and voucher threats improve low-performing schools? *Journal of Public Economics, 90 (*2006), 239-255.

Fuhrman, S. (2004). Introduction. In S. Fuhrman and R. Elmore (Eds.), *Redesigning accountability systems for education*. New York: Teachers College Press.

Garfinkel, I. & Gramlich, E. (1973). A statistical analysis of the OEO experiment in educational performance contracting. *The Journal of Human Resources, 8*(3), 275-305.

Gelman, A., & Imbens, G. (2014). Why high-order polynomials should not be used in regression discontinuity designs. NBER Working Paper series. Paper No. 20405.

Gilraine, M. (2016). School accountability and the dynamics of human capital formation. URL: http://tinyurl.com/Gilraine-JMP

Gramlich, E., & Koshel, P. (1975). *Educational performance contracting: An evaluation of an experiment.* Washington D.C.: The Brookings Institution

Gronberg, T. J. & Jansen, D. M. (2006). Introduction. In T. J. Gronberg and D. W. Jansen (Eds.) *Advances in Applied Microeconomics (Vol 14), Improving school accountability: Check-ups or choice.* Amsterdam: Elsevier.

Hannaway, J. (1996). Management decentralization and performance-based incentives: theoretical consideration for schools. In E. Hanushek and W. Jorgenson (Eds.) *Improving America's schools: the role of incentives.* United States of America: National Academy of Science.

Hanushek, E. (1996). Outcome, costs and incentives in schools. In E. Hanushek and D. Jorgenson (Eds.), *Improving America's schools: the role of incentives.* United States of America: National Academy of Sciences.

Hanushek, E., Benson, C., Freeman, R, Jamison, D., Levin, H., Maynard, R., Murnane, R., Rivkin, S., Sabot, R., Solmon, L., Summers, A., Welch, F., & Wolfe, B. (1994). *Making schools work: improving performance and controlling costs.* Washington, D.C.: The Brookings Institution.

Hanushek, E., & Raymond, M. (2005). Does school accountability lead to improved school performance? *Journal of Policy Analysis and Management, 24* (2), 297-327.

Hart, C., & Figlio, D. (2015). School accountability and school choice: Effects of student selection across schools. *National Tax Journal, 68* (35), 875-900.

Hillman, N. (2016). Why performance-based college funding doesn't work. The Century Foundation. Retrieved Sept. 18th, 2016 from: http://tcf.org/content/report/why-perofrmnce-based-college-funding-doesnt-work/

Holland, P. (1986). Statistics and causal inference. *Journal of American Statistical Association, 81*(396), 945-960.

Holmstrom, B., & Milgrom, P. (1994). The firm as an incentive system. *The American Economic Review, 84* (4), 972-991.

Hussain, I. (2015). Subjective performance evaluation in the public sector evidence from school inspections. *Journal of Human Resources, 50*(1), 189-221.

Imbens, G. W., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of econometrics, 142(*2), 615-635.

Imbens, G. W., & Kalyanaraman, K. (2012). Optimal Bandwidth Choice for the Regression Discontinuity Estimator. *The Review of Economic Studies, 79(*3), 933-959.

Jackson, K., Johnson, R., & Persico, C. (2015). The effects of school spending on educational and economic outcomes: Evidence from school finance reforms. *Institute for Policy Research Northwestern University, Working Paper Series, WP-15-19*, 1-83.

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

Jacob, B. A. (2007). Test-based accountability and student achievement: an investigation of differential performance on NAEP and state assessments. Working paper 12817. *National Bureau of Economic Research*.

Jacob, R. T., Zhu, P., Somers, M. A., & Bloom, H. S. (2012). A practical guide to regression discontinuity (pp. 1-91). New York: MDRC.

Kane, T., & Staiger, D. (2002). The promise and pitfalls of using imprecise school accountability measures. *Journal of Economic Perspectives, 16* (4), 91-114.

Kahneman, D., & Tversky, A. (1984). Choices, values, frames. *American Psychologist, 39*(4), 341-350.

Klein, S., Hamilton, L., McCaffrey, & Stecher, B. (2000). What do test scores in Texas tell us? *Education Policy Analysis Archives, 8* (49), 1-22.

Koretz, D. (2005). Alignment, High Stakes, and the Inflation of Test Scores. CSE Report 655. *National Center for Research on Evaluation, Standards, and Student Testing (CRESST)*.

Kubanek, J., Snyder, L., Abrams, R. (2015). Reward and punishment act as distinct factors guiding behavior. *Cognition, 139*(2015), 154-167.

Ladd, H. (1996). Introduction. In H. Ladd (Ed.), *Holding schools accountable: performance-based reform in education*. Washington, D.C.: The Brookings Institution.

Laffont, J-J., & Martimort, D. (2002). The theory of incentives: The principal-agent model. New Jersey: Princeton University Press.

Lafortune, J., Rothstein, J., & Schanzenbach, D. (2016). School finance reform and the distribution of student achievement. *NBER Working Paper Series.* Working paper N 22011.

Lee, J. (2008). Is test-driven external accountability effective? Synthesizing the evidence from cross-state causal-comparative and correlational studies. *Review of Educational Research, 78*(3), 608-644.

Lee, D. S., & Lemieux, T. (2009). Regression discontinuity designs in economics. *NBER Working Paper Series.* Working paper N 14723.

Leuven, E., Lindahl, M., Oosterbeek, H., & Webbink, D. (2007). The effect of extra funding for disadvantaged pupils on achievement. *The Review of Economics and Statistics, 89* (4), 721-736.

Levin, H.M. (1974). A conceptual framework for accountability in education. *The School Review, 82* (3), 363-391.

Levin, H. M. (1980). Educational Production Theory and Teacher Inputs. In Bidwell and Windham (Eds.), *The Analysis of Educational Productivity*. Cambridge, MA: Ballinger.

Ley, Decreto con fuerza de Ley 2 (August 1998). Fija texto refundido, coordinado y sistematizado del decreto con fuerza de ley 2, de 1996, sobre subvención del estado a establecimientos educacionales. Biblioteca Congreso Nacional: Chile

Ley, Decreto, 235 (April 2008). Aprueba reglamento de la ley 20.248, que establece una subvención escolar preferencial para niños y niñas prioritarios. Biblioteca Congreso Nacional: Chile

Ley, Decreto, 293(August 2009). Establece estándares nacionales y criterios específicos para la calificación de los resultados educativos a que se refieren los artículos 9 y 10 de la ley 20.248. Biblioteca Congreso Nacional: Chile

Ley, N. 20.248 (January 2008). Ley de Subvención Escolar Preferencial. Biblioteca Congreso Nacional: Chile

Ley, N. 20.501 (February 2011). Calidad y Equidad de la educación. Biblioteca Congreso Nacional: Chile

Ley, N. 20.550 (August 2011). Modifica Ley de Subvención Escolar Preferencial. Biblioteca Congreso Nacional: Chile

Ley, N. 20.529 (August 2011). Sistema nacional de aseguramiento de la calidad de la educación parvularia, básica y media y su fiscalización. Biblioteca Congreso Nacional: Chile

Linn, R. L. (2004). Accountability models. In S. Fuhrman and R. Elmore (Eds.), *Redesigning* accountability systems for education. New York: Teachers College Press.

Ludwig, J., Kling, J. R., & Mullainathan, S. (2011). Mechanism experiments and policy evaluations. *The Journal of Economic Perspectives, 25*(3), 17-38.

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

MINEDUC (2012). Impacto de la Ley SEP en SIMCE: una Mirada a 4 años de su implementación. Serie *Evidencias, 1* (8).

Mintrop, H., & Sunderman, G. (2009). Predictable failure of federal sanctions-driven accountability for school improvement and why we may retain it anyway. *Educational Researcher,38*(5), 353-364.

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

Mizala, A., & Torche, F. (2013). ¿Logra la subvención escolar preferencial igualar los resultados educativos? Espacio Público. Documento de referencia (9).

Mizala, A., & Urquiola, M. (2013). School markets: The impact of information approximating schools' effectiveness. *Journal of Development Economics, 103*(2013), 313-335.

National Research Council. (2011). *Incentives and test-based accountability in education.* Committee on Incentives and Test-based Accountability in Public Education, M. Hout and S. W. Elliot, editors. Board of Testing and Assessment, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Navarro-Palau, P. (2015). Effects of differentiated school vouchers: Evidence from a policy change and date of birth cutoffs. Dissertation submitted in partial fulfilment of the requirements of the degree of doctor of philosophy to the Graduate School of Arts and Sciences, Columbia University.

Neal, D. & Schanzenbach, W. (2010). Left behind by design: proficiency counts and test-based accountability. *The Review of Economics and Statistics, 92* (2), 263-283.

Neilson, C. (2013). Targeted vouchers, competition among schools, and the academic achievement of poor students. Yale University. URL: http://pantheon.yale.edu/~can7/Neilson_2013_JMP_current.pdf (accessed 11.12.2014).

O'Day, J. (2004). Complexity, accountability, and school improvement. In S. Fuhrman and R. Elmore (Eds.), *Redesigning accountability systems for education*. New York: Teachers College Press.

OECD (2013). PISA 2012 Results: What makes schools successful? Resources, policies and practice (Volume IV). Paris: PISA, *OECD Publishing*.

Papay, J. P., Willett, J. B., & Murnane, R. J. (2011). Extending the regression-discontinuity approach to multiple assignment variables. *Journal of Econometrics, 161*(2), 203-207.

Papay, J. P., Murnane, R. J., & Willett, J. B. (2014). High-School Exit Examinations and the Schooling Decisions of Teenagers: Evidence From Regression-Discontinuity Approaches. *Journal of research on educational effectiveness, 7(*1), 1-27.

Peirano, C. & Vargas, J. (2005). Private schools with public financing in Chile. In L. Wolff, J.C. Navarro and P. Gonzalez (eds), Private education and public policy in Latin America. Washington D.C.: PREAL.

Peterson, G. (1974). The distributional impact of performance contracting in schools. In H. Hochman and G. Peterson (Eds.), Redistribution through public choice. New York: Columbia University Press.

Podgursky, M., & Springer, M. (2007). Teacher performance pay: A review. Journal of *Policy Analysis and Management, 26* (4), 909-950.

Rau, T., & Contreras, D. (2009). Tournaments, gift exchanges, and the effect of monetary incentives for teachers: The case of Chile. Unpublished mimeo. Retrieved fro m http://www2. econ. iastate. edu/faculty/Orazem/TPS papers/Contreras. pdf.

Reardon, S. F., & Robinson, J. P. (2012). Regression discontinuity designs with multiple rating-score variables. *Journal of Research on Educational Effectiveness, 5(*1), 83-104.

Richards, C., & Sheu, T. M. (1992). The South Carolina school incentive reward program: A policy analysis. *Economics of Education Review, 11* (1), 71-86.

Robinson, J. (2008). Essays on the effectiveness of policies and practices for reducing cognitive gaps between linguistic groups and socioeconomic groups. Dissertation submitted in partial fulfilment of the requirements of the degree of doctor of philosophy to the School of Education and the Committee on Graduate Studies of Stanford University.

Rockoff, J., & Turner, L. (2008). Short run impacts of accountability on school quality. *NBER Working Paper Series*. Working paper 14564. URL (Retrieved July 20, 2015): http://www.nber.org/papers/w14564

Rouse, C., Hannaway, E., 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., & Deci, E. (1999). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology,* 25, 54-67.

Sahlberg, P. (2010). Rethinking accountability in a knowledge society. *Journal of Educational Change, 11*, 45-61.

Schochet, P., Cook, T., Deke, J., Imbens, G., Lockwood, J.R., Porter J. & Smith, J. (2010). Standards for Regression Discontinuity Designs. Retrieved from What Works Clearinghouse website: http://ies.ed.gov/ncee/wwc/pdf/wwc_rd.pdf.

Schochet, P. (2009). Statistical power for regression discontinuity designs in education evaluations. *Journal of Educational and Behavioral Statistics, 34* (2), 238-266.

Scott-Clayton, J. (2008). On money and motivation: A quasi-experimental analysis of financial incentives for college achievement. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.187.644&rep=rep1&type=pdf

Skovron, C., & Titiunik, R. (2015). A Practical Guide to Regression Discontinuity Designs in Political Science. working paper, University of Michigan.

Social Finance (2011). A technical guide to commissioning social impact bonds. Retrieved June 2014: www.socialfinance.org.uk

Thorndike, E. (1927). The law of effect. *The American Journal of Psychology*, *39*(1/4), 212-222.

Tokman, A. (2002). Evaluation of the P900 program: A targeted education program for underperforming schools. Working paper 170. *Banco Central de Chile.*

Tsang, M., & Levin, H.M. (1983). The impact of intergovernmental grants on educational expenditure. *Review of Educational Research, 53* (3), 329-367.

Urquiola, M., & Verhoogen, E. (2009). Class-size caps, sorting and the regression discontinuity design. *American Economic Review, 99*(1), 179-215.

Villarroel, G. (2012). Mejoramiento en resultados académicos de la educación básica en Chile: Primeros efectos de la ley de Subvención Escolar Preferencial (SEP). Tesis para optar al grado de Magister en Economía, Universidad de Chile.

Witte, J., Wolf, P., Cowen, J., Carlson, D., & Fleming, D. (2014). High stakes choice: Achievement and accountability in the Nation's oldest urban voucher program. *Educational Evaluation and Policy Analysis, 36* (4), 437-456.

Weiner, J., Donaldson, M., & Dougherty, S. M. (2016). Missing the Boat–Impact of Just Missing Identification as a High-Performing School. *Leadership and Policy in Schools*, 1-26.

Woessmann, L., Ludemann, E., Schutz, G., and West, M. (2007). School accountability, autonomy, choice and the level of achievement: International evidence form PISA 2003. *OECD Education Working Papers, 13*. OECD Publishing.

Wooldridge J. (2002). *Econometric analysis of cross section and panel data.* Cambridge, MA: Massachusetts Institute of Technology.

Yechiam, E., & Hochman, G. (2013). Losses as modulators of attention: Review and analysis of the unique effects of losses over gains. *Psychological Bulletin, 139* (2), 497-518.

Zubizarreta, J., Paredes, R., & Rosenbaum, P. (2014). Matching for balance, pairing for heterogeneity in an observational study of the effectiveness of for-profit and not-for-profit high schools in Chile. *The Annals of Applied Statistics, 8* (1), 204-231.

Appendix A. School classification process

According to the Decree 293 and the Technical Report of *Proceso de Clasificacion SEP*, the process to classify schools is as follows.

Data needed:

a. School ID
b. Type of school
c. Average number of students who take SIMCE test for the last three years available (average of math, language and sciences).[82]
d. Average SIMCE scores for last three years available (average of math, language and sciences)
e. Proportion of students who score above 250 points on SIMCE for the last three years available (average of math, language and sciences).
f. Proportion of students who score above 300 points on SIMCE for the last three years available (average of math, language and sciences).
g. Socioeconomic group in which the school is classified according to data from SIMCE for the last three years.
h. Approval rate for last year.[83]
i. Retention rate for last year.
j. SNED Improvement index for last year.
k. SNED Integration index for last year.
l. SNED Initiative index for last year.
m. Teacher evaluation scores for last year.

Step 0: Prepare data

i. Calculate median of average SIMCE scores for each socioeconomic group using all the schools.
ii. Calculate median of proportion of students scoring above 250 points on SIMCE for each socioeconomic group using all the schools.
iii. Calculate median of proportion of students scoring above 300 points on SIMCE for each socioeconomic group using all the schools.
iv. Standardize all the complementary indicators (approval rate, retention rate, improvement, initiative, integration, teacher evaluation) using the data for all the schools.
v. Standardize average SIMCE score for years available in the last three years.

Step 1: Check which schools can be classified according to the classification formulae.

[82] The last three years of available SIMCE scores skips the year prior to the one of the classification. For example, if the classification of schools for year 2012 is done in 2011, then the SIMCE scores considered are from years 2010, 2009 and 2008.

[83] Approval rates, retention rates, SNED improvement, integration and initiative indices and Teacher evaluation scores are all available for the year prior to the classification year. For example, if the classification of schools for year 2012 is done in 2011, then these indices are from year 2011 (reported numbers from the previous year).

i. Does the school have at least two SIMCE measure in the last three years? If the school does not have at least two measures, then the school is classified as "emergent".
ii. Does the school have less than 20 students taking the test on the average of the last 3 years of SIMCE? If the school has an average of students taking the test of 20 or lower, then the school is classified as "emergent".
iii. If the school has two or more SIMCE measures AND has an average of students taking the test greater than 20, then the school can be classified using the formula.

Step 2: Apply classification formula only for schools that can be classified.

i. Does the average SIMCE score per year is less than 220 points?
ii. Does the average SIMCE score per year is greater than the median of the socioeconomic group?
iii. Does the proportion of students scoring above 250 points is less than 20%?
iv. Does the proportion of students scoring above 250 points is greater than the median of the socioeconomic group?
v. Does the proportion of students scoring above 300 points is greater than the median of the socioeconomic group?
vi. Does rule i and iii happen simultaneously in at least 2 years of SIMCE scores? If yes, then school is preliminarily classified as "in recovery".
vii. Do rules ii, iv and v happen simultaneously in at least 2 years of SIMCE scores? If yes, then school is preliminarily classified as "autonomous".
viii. Schools that are not classified as "in recovery" by rule vi or "autonomous" by rule vii, are preliminarily classified as "emergent".
ix. Construct Complementary Index of School Quality (IIC).
    a. For public schools: Weighted average of standardized quality indices (Approval rate* 0.25, Retention rate*0.25, Improvement*0.17, Initiative*0.13, Integration*0.13, Teacher evaluation*0.07)
    b. For private voucher schools: Weighted average of standardized quality indices (Approval rate* 0.25, Retention rate*0.25, Improvement*0.20, Initiative*0.15, Integration*0.15)
x. Construct Index of Educational Quality (ICE): Weighted average of standardized average of SIMCE score * 0.7 and IIC*0.3
xi. Calculate the 10^{th} percentile of the ICE scores.
xii. Calculate the median of the ICE scores for each socioeconomic group only for schools that can be classified.
xiii. Does the school has an ICE score lower than the 10^{th} percentile? If yes, then a school preliminarily classified as "emergent" becomes "in recovery".
xiv. Does the school has an ICE score less than the median of the socioeconomic group? If yes, then a school preliminarily classified as "autonomous" becomes "emergent".

Appendix B. SIMCE evaluation calendar

Table B.1. SIMCE evaluation calendar.

Grade	Subject	Years							
		2008	2009	2010	2011	2012	2013	2014	2015
4th grade	Math	X	X	X	X	X	X	X	X
	Language	X	X	X	X	X	X	X	X
	Social sciences	X		X		X		X	
	Natural sciences		X		X		X		X
8th grade	Math		X		X		X	X	X
	Language		X		X		X	X	X
	Social sciences		X		X			X	
	Natural sciences		X		X		X		X

Appendix C. Additional tables and figures

Table C.1. Coincidence of school classification between MINEDUC and author's calculations (percentage).

	Years			
School classification	2012	2013	2014	2015
In recovery	92.15	95.43	98.21	100
Emergent	89.23	89.7	83.59	84.48
Autonomous	96.44	94.96	95.44	94.98
Total coincidence	92.07	92.01	88.61	89.06

Note: I have calculated which category corresponds to each school according to the rules specified in the law, the decree and the technical document. Each percentage specified here indicates that the classifications I have made correspond to the ones defined by MINEDUC. E.g., from all schools MINEDUC classified as emergent on 2014, I can match the classification using my interpretation of the formula in 83.59% of the cases.

Table C.2. First stage models for schools being classified or not as "autonomous". Year 2013 and pooled rounds.

	(1)	(2)	(3)	(4)
	2013 round		Pooled rounds: 2012 & 2013	
Above ICE $_{t-1}$	0.32***	0.31***	0.31***	0.31***
	(0.03)	(0.03)	(0.02)	(0.02)
ICE $_{t-1}$	0.01	0.02	0.01	0.00
	(0.01)	(0.01)	(0.01)	(0.00)
Above ICE $_{t-1}$ * ICE $_{t-1}$	1.11***	1.11***	1.22***	1.22***
	(0.05)	(0.05)	(0.04)	(0.04)
Controls	No	Yes	No	Yes
Observations	2396	2396	4620	4620
F	1996.90	1535.63	4736.13	4102.17
Bandwidth	0.857	0.857	0.779	0.779

Notes: The table shows resulting coefficients from equation 5 estimated with a linear probability model. The cells contain the main coefficients and the robust standard errors on parenthesis. The schools included are those SEP schools that are classified according to the classification formula. Covariates included in columns (2) and (4) control for whether the school is in an urban area, the type of school, schools SES level, and the enrollment rates in 4th grade. Columns (3) and (4) include a year fixed effect to account per any cohort trend.
* p<0.1, **p<0.05, ***p<0.01

Table C.3. Impact of "autonomous" classification on 2013 round on high- and low-stakes grades.

	(1)	(2)	(3)	(4)
	All schools		Private voucher schools by profit status	
Outcomes	Baseline	With controls	For-profit	Not for-profit
Panel A: High stakes outcomes				
4th grade Math scores on year t	6.29*	1.90	2.03	3.60
	(2.78)	(2.66)	(4.39)	(7.95)
	2396	2396	619	419
4th grade Language scores on year t	3.08	-1.13	-3.57	2.60
	(2.57)	(2.30)	(3.94)	(5.99)
	2396	2396	619	419
4th grade Science scores on year t	3.71	-0.78	-2.42	1.63
	(2.23)	(1.96)	(3.61)	(4.70)
	2392	2392	619	418
Panel B: Low stakes outcomes				
8th grade Math scores t	7.98**	1.71	-5.55	7.65
	(2.49)	(2.04)	(4.11)	(6.22)
	2346	2346	590	411
8th grade Language scores t	3.21	-1.80	-8.64	4.60
	(2.62)	(2.32)	(4.65)	(6.17)
	2344	2344	591	411
8th grade Science scores t	7.10*	0.74	-2.00	-1.70
	(2.53)	(2.08)	(4.26)	(5.81)
	2345	2345	588	411

Note: The table shows resulting coefficients from equation 6 with covariates. The cells contain the main coefficient, the robust standard error on parenthesis, and the sample size. Models are estimated using OLS. The schools included in the sample are those SEP schools that are classified using the classification formula, and that are within a 0.857 bandwidth. The two columns at the right only consider private voucher schools, and divides the sample on whether the school are for-profit or not.
* p<0.1, **p<0.05, ***p<0.01

Table C.4. Impact of "autonomous" classification on pooled rounds of 2012 and 2013 on high-stake outcomes.

Outcomes	Pooled rounds: 2012 & 2013
4th grade Math scores on year t	1.69
	(1.88)
	4620
4th grade Language scores on year t	0.18
	(1.62)
	4618
4th grade Science scores on year t	0.45
	(1.47)
	4616

***Note*:** The table shows resulting coefficients from equation 6 with covariates and year fixed effect to account per any cohort trend on SIMCE scores. The cells contain the main coefficient, the robust standard error on parenthesis, and the sample size. Models are estimated using OLS. The schools included in the sample are those SEP schools that are classified using the classification formula, and that are within a 0.779 bandwidth.
* p<0.1, **p<0.05, ***p<0.01

Table C.5. Impact of "autonomous" classification on 2012 on high-stakes outcomes using other IVs.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	SIMCE	SIMCE	SIMCE	P250	P250	P250	P300	P300	P300
Outcomes	t-2	t-3	t-4	t-2	t-3	t-4	t-2	t-3	t-4
4th grade Math scores on year t	11.31	-0.14	11.47	-2.40	-3.77	24.16**	1.59	11.96	9.27
	(7.49)	(5.83)	(7.25)	(9.83)	(8.04)	(7.66)	(9.14)	(9.10)	(10.50)
	1983	1871	1942	1957	1818	1927	1964	2000	2138
4th grade Math scores on year t+1	-0.08	3.17	6.68	3.73	-11.02	27.48***	9.94	-8.86	17.11
	(7.17)	(5.87)	(7.59)	(9.45)	(8.13)	(7.70)	(8.85)	(10.16)	(10.24)
	1968	1855	1925	1941	1807	1911	1948	1983	2117
4th grade Language scores on year t	7.96	5.30	10.36	-4.18	4.18	21.34**	1.88	10.32	7.58
	(6.08)	(4.93)	(6.67)	(8.56)	(6.79)	(6.99)	(7.81)	(7.88)	(9.50)
	1981	1869	1940	1955	1817	1925	1962	1998	2136
4th grade Language scores on year t+1	-4.63	1.11	-1.60	6.17	-1.63	14.76*	-2.28	2.56	11.85
	(6.11)	(5.17)	(6.57)	(7.87)	(6.44)	(6.74)	(8.08)	(8.23)	(8.96)
	1968	1855	1925	1941	1807	1911	1948	1983	2117
4th grade Science scores on year t	11.63	0.16	10.86	-3.72	2.56	21.61**	-1.22	14.72	15.58
	(6.17)	(5.17)	(6.09)	(8.05)	(6.57)	(6.29)	(7.54)	(7.55)	(8.57)
	1983	1871	1942	1957	1818	1927	1964	2000	2138
4th grade Science scores on year t+1	-1.21	3.78	3.78	3.82	-4.80	14.79*	-0.18	0.89	14.56
	(5.42)	(4.53)	(5.91)	(6.96)	(6.19)	(5.86)	(6.88)	(7.63)	(8.01)
	1961	1847	1919	1934	1800	1905	1938	1974	2110
Covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Bandwidth	1.175	1.152	1.042	1.191	1.071	1.238	.921	.975	1.256

Note: The table shows resulting coefficient B_1 from equation 6 with covariates, but with the instrumental variables specified on the top of the column. The cells contain the main coefficient, the robust standard error on parenthesis, and the sample size. All outcomes are estimated using OLS. The schools included in the sample are those SEP schools that are classified using the classification formula, and that are within the bandwidth specified on the bottom row. At the end of each column there is the specified bandwidth used for the analysis of all the outcomes of the column. This bandwidth is the optimal bandwidth following Calonico, Cattaneo and Titiunik (2014) for the 4th grade math scores on year t. I have used the same bandwidth for all the outcomes for convenience. However the bandwidths estimated for the other outcomes do not vary much.

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

Table C.6. Minimum detectable effects for the threshold of the "autonomous" classification in 2012.

Outcome	MDE	MDES
Panel A: High stakes outcomes		
4th grade Math scores on year t	7.73	0.15
4th grade Math scores on year t+1	7.92	0.16
4th grade Language scores on year t	6.68	0.13
4th grade Language scores on year t+1	6.68	0.13
4th grade Science scores on year t	6.28	0.13
4th grade Science scores on year t+1	6.08	0.12
Panel B: Low stakes outcomes		
8th grade Math scores on year t+1	7.47	0.15
8th grade Language scores on year t+1	7.50	0.15
8th grade Science scores on year t+1	7.13	0.14

Note: Following Schochet (2009), I calculate the minimum detectable effect (MDE) values using the formula

$$MDE = Factor\ (\propto, \beta, df) * \sqrt{Var(impact)}$$

Where, $Var(impact)$ is the variance of the impact estimate controlling for covariates (Table 25, column 2; Table 26, column 2), and $Factor\ (.)$ is a constant that is a function of the significance level ($\propto$), statistical power (β), and the number of degrees of freedom (df). I use a significance level of 0.05, a power of 0.80, for a two-tailed test, as it is commonly used. The constant is 2.83. I also report MDES which is the MDE in effect size units, i.e., as a percentage of the standard deviation of the outcome measure. $MDES = MDE / \sigma$, where σ is the standard deviation of the outcome measure for all the population.

Table C.7. Minimum detectable effects for the threshold of the "in recovery" classification in 2012 and 2013 rounds pooled.

Outcome	MDE	MDES
Panel A: High stakes outcomes		
4th grade Math scores on year t	7.92	0.16
4th grade Math scores on year t+1#	12.17	0.24
4th grade Language scores on year t	7.16	0.14
4th grade Language scores on year t+1#	10.56	0.21
4th grade Science scores on year t	6.45	0.13
4th grade Science scores on year t+1#	8.57	0.17
Panel B: Low stakes outcomes		
8th grade Math scores on year t*	6.96	0.14
8th grade Math scores on year t+1#	8.32	0.17
8th grade Language scores on year t*	6.23	0.12
8th grade Language scores on year t+1#	7.61	0.15
8th grade Science scores on year t*	10.33	0.21
8th grade Science scores on year t+1#	7.10	0.14

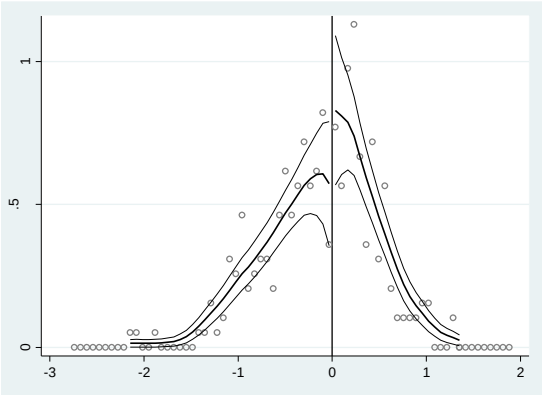
Note: Following Schochet (2009), I calculate the minimum detectable effect (MDE) values using the formula

$$MDE = Factor\ (\propto, \beta, df) * \sqrt{Var(impact)}$$

Where, $Var(impact)$ is the variance of the impact estimate controlling for covariates (Table 33, column 2; Table 34 columns 2 and 4), and $Factor\ (.)$ is a constant that is a function of the significance level ($\propto$), statistical power (β), and the number of degrees of freedom (df). I use a significance level of 0.05, a power of 0.80, for a two-tailed test, as it is commonly used. The constant is 2.83. I also report MDES which is the MDE in effect size units, i.e., as a percentage of the standard deviation of the outcome measure. $MDES = MDE / \sigma$, where σ is the standard deviation of the outcome measure for all the population.

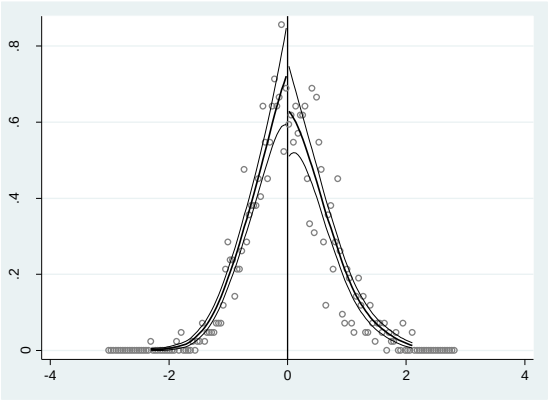
#Outcome for round of 2012 only. *Outcome only available for round of 2013.

Panel A. Low SES schools



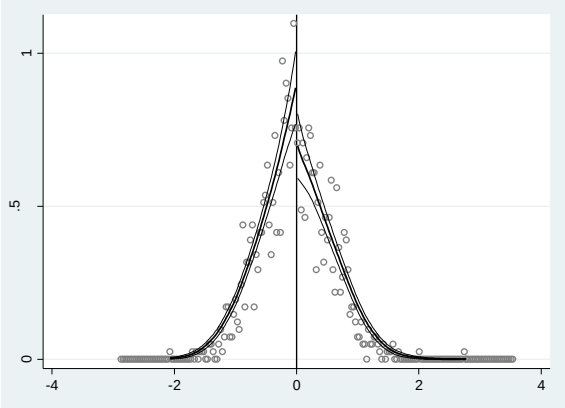
Log difference in height= 0.400; *p*=.29; n=295

Panel C. Mid SES schools



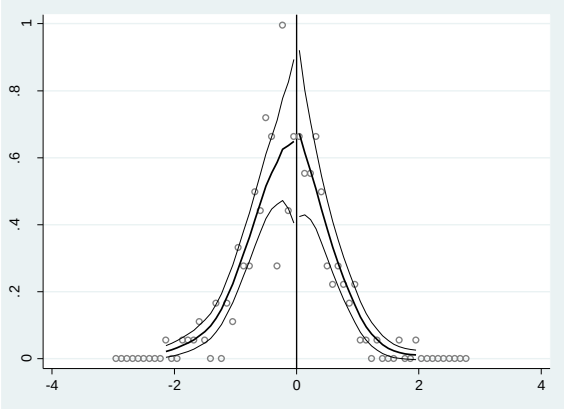
Log difference in height=-0.145; *p*=.14; n=1069

Panel B. Mid Low SES schools



Log difference in height=-0.245; *p*=.11; n=1338

Panel D. Mid-High and High SES schools



Log difference in height=0.076; *p*=.30; n=199

Figure C.1. McCrary density test. Density of ICE of schools by SES group close to the "autonomous" cutoff.

Note: Schools included are all those elementary schools that have adopted the SEP law, that are classified using the classification formulae. Data from 2012.

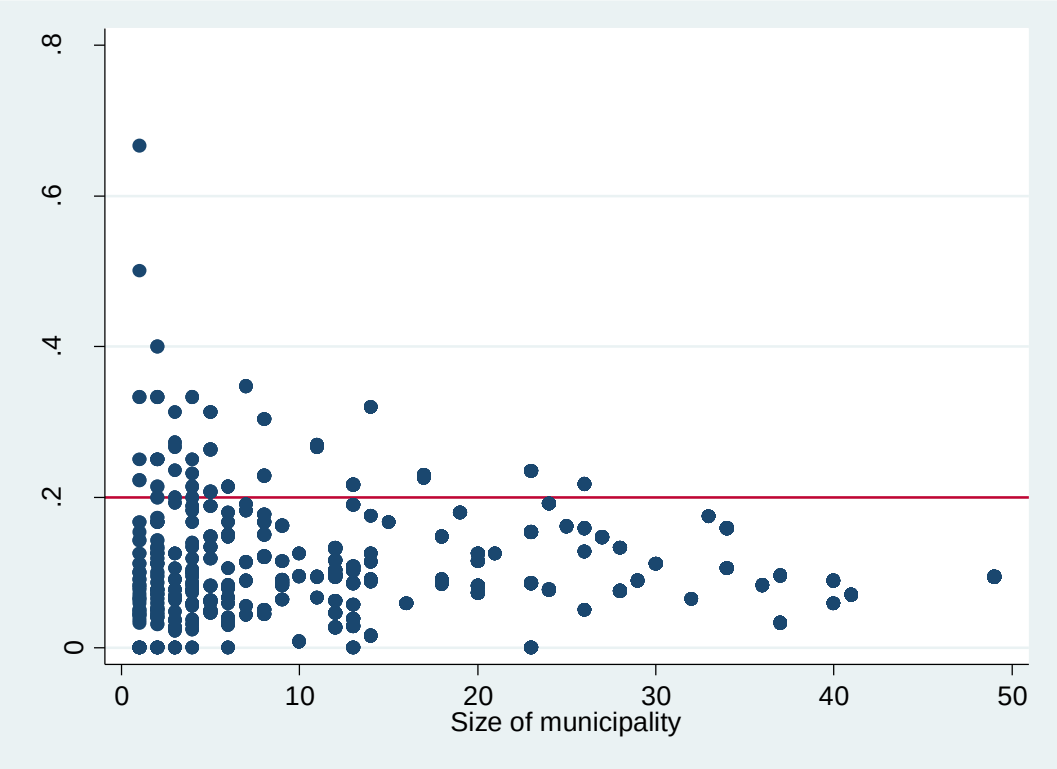


Figure C.2. Share of "autonomous" schools by municipality size.

Note: Schools included in the sample are those SEP schools that are classified using the classification formula, and that are within the 0.706 bandwidth.

This page intentionally left blank